ML based Chunking

Karan Agarwalla	Rishi Agarwal	Sumanyu Ghoshal
180050045	180050086	180070060

Abstract

Text chunking refers to the task of splitting a string of textual information into non-overlapping groups of related units. We explore three popular techniques for this task. The proposed techniques are applied to the CONLL 2020 corpus. We perform ablation tests to find the importance of POS tags in chunking. The results in all three models indicate that POS tags play an important role in chunking.

Techniques used for chunking

In this section, we describe our models - MEMM, CRF and Bi-LSTM and implementation details. Our evaluation metrics are precision, recall and f1-score. The results for each model are shown as well.

1. MEMM

MEMM(Maximum-Entropy Markov Model) is a graphical model for sequence labelling that combines features of Hidden Markov Models(HMMs) and Maximum Entropy(MaxEnt) models. 'Improved Iterative Scaling'(**IIS**) algorithm has been used to train the model. Further max_iter has been set to 8 to train for both the cases.

We use **morphological** features (like word prefixes and suffixes), **POS tags** and capitalisation features. We use a window size of 6 for POS tags and the previous Chunk Label.

We performed an **ablation test** to find the importance of POS tags in the feature vector. Below are two tables showing the performance of the model in the two cases. We find that the model is significantly better when POS tags are a part of the feature vector.

	Precision	Recall	F1 score
B tag	0.93	0.91	0.92
I tag	0.87	0.89	0.88
Overall (weighted average)	0.91	0.91	0.91

Table 1.1: The precision, recall , F1-score values are reported here for the MEMM model when POS tags are **not** included in the feature vector

	Precision	Recall	F1 score
B tag	0.94	0.93	0.93
I tag	0.90	0.90	0.90
Overall (weighted average)	0.92	0.92	0.92

Table 1.2: The precision, recall, F1-score values are reported when both word and POS tag have been taken

2. CRF

Conditional random fields are a **log linear model** for sequential tasks. We use the popular '**Limited memory BFGS'** algorithm to find the parameters of the model. In addition, we use **elastic net** regularization which linearly combines L1 and L2 penalties. We set the max iterations to 100.

We use **morphological** features (like word prefixes), **POS tags** and other features indicating if the word is **uppercase** or lowercase, contains **digit** or not, etc. We use a window size of 2 while considering neighbouring words. We tried experimenting with the window size and found that increasing window size from 1 to 2 improves the performance of the model.

We performed an **ablation test** to find the importance of POS tags in the feature vector. Below are two tables showing the performance of the model in the two cases. We find that the model is significantly better when POS tags are a part of the feature vector.

	Precision	Recall	F1 score
B tag	0.95	0.94	0.95
I tag	0.91	0.93	0.92
Overall (weighted average)	0.94	0.94	0.94

Table 2.1: The precision, recall, f1-score values are reported here for the CRF model when POS tags are **not** included in feature vector

	Precision	Recall	F1 score
B tag	0.96	0.96	0.96
I tag	0.94	0.94	0.94
Overall (weighted average)	0.95	0.95	0.95

Table 2.2: The precision, recall, f1-score values are reported here for the CRF model when POS tags are included in feature vector

3. Bi LSTM

	Precision	Recall	F1 score
B tag	0.95	0.93	0.94
I tag	0.90	0.92	0.91
Overall (weighted average)	0.93	0.93	0.93

Table 3.1: The precision, recall, f1-score values are reported here for the Bi-LSTM when POS tags are **not** included

	Precision	Recall	F1 score
B tag	0.93	0.96	0.94
I tag	0.93	0.88	0.91
Overall (weighted average)	0.93	0.93	0.93

Table 3.2: The precision, recall, f1-score values are reported when both word and POS tag have been taken

	Precision	Recall	F1 score
B tag	0.96	0.97	0.97
I tag	0.96	0.94	0.95
Overall (weighted average)	0.96	0.96	0.96

Table 3.3: The precision, recall, f1-score values are reported when only POS tag has been taken

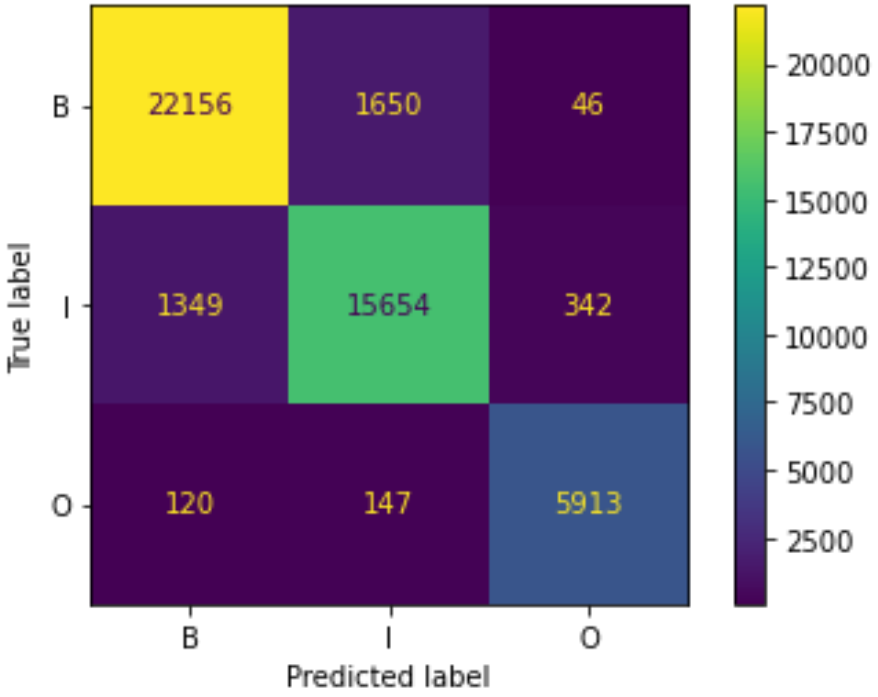
MEMM

1. Used MaxEnt Classifier with the following **Feature Set**:
   - Prefix features of current and previous word(First two and three letters)
   - Suffix features of current and previous word(Last two and three letters)
   - Capitalisation of current and previous word
   - POS tags for previous three words, current word and next two words
   - Previous Chunk Tag
   - Begin of Sentence['BOS'] and end of sentence ['EOS']
   - A bias term
2. Used Viterbi Decoding to obtain the correct tag sequence

Some of the Informative Features of Classifier**:**

- word[:2]=="of" and label is 'O'
- EOS==True and label is 'B'
- word[:2]=="in" and label is 'O'
- -1:word[-3:]=="for" label is 'I'
- postag=="PRP" and label is 'O'
- postag=="NNS" and label is 'O'

Confusion Matrix



Classification Report:

	precision	recall	f1-score	support
B	0.94	0.93	0.93	23852
I	0.90	0.90	0.90	17345
O	0.94	0.96	0.95	6180
accuracy			0.92	47377
macro avg	0.92	0.93	0.93	47377
weighted avg	0.92	0.92	0.92	47377

Abalitive Study:

Window size of 6 for POS Tags was optimal in case of the given feature set as it provided better results

Results after removal of POS Tags:

	precision	recall	f1-score	support
B	0.93	0.91	0.92	23852
I	0.87	0.89	0.88	17345
O	0.94	0.94	0.94	6180
accuracy			0.91	47377
macro avg	0.91	0.91	0.91	47377
weighted avg	0.91	0.91	0.91	47377

We observe a better performance on including POS Tags than removing them. However, on carefully observing the most important features upon removal of POS Tags(they are prefixes like “hey” and “fr”), one realises that prediction is somewhat arbitrary in nature and the model has not learnt important features.
Also all metrics are better on including POS tags.

Misclassification examples:

Presence of POS Tags

Under the existing contract , Rockwell said , it has already delivered 793 of the shipsets to Boeing .
True Tag Sequence and Predicted Tag Sequence with POS Tags:
['B','B','I','I','O','B','B','O','B','B','I','I','B','B','B','I','B','B','O']
Predicted Tag Sequence without POS Tags:
['B','B','I','I','O','B','B','O','B','B','I','I','I','B','B','I','B','I','O']

We observe that excluding POS Tags results in misclassification of “Boeing”. This can be attributed to the fact the feature representation of “Boeing” without POS Tags is arbitrary. Moreover the suffix “ing” is a verb suffix resulting in further misinterpretation. The inclusion of POS tags provides some structure to the word and the model also has access to nearby tags in the sentence.

Fragmentation

Manville is a building and forest products concern .
True Tag Sequence:
['B','B','B','I','I','I','I','I','O']
Predicted Tag Sequence:
['B','B','B','I','O','B','I','B','O']

The above tag sequence is correctly predicted by Bi-LSTM and CRF models.

He is now changing the place he sleeps every night , sometimes more than once a night .
True Tag Sequence:
['B','B','I','I','B','I','B','B','B','I','O','B','I','I','I',’B','I','O']
Predicted Tag Sequence:
['B','B','B','B','B','I','B','B','B','I','O','B','B','B','B','B','I','O']

We observe fragmentation in case of MEMM modelling. This is because MEMM considers transitions from current state and hence has a local scope while considering transitions.

Other Examples

Net income : $ 44.9 million ; or 25 cents a share
True Tag Sequence:
['B','I','O','B','I','I','O','O','B','I','B','I']
Predicted Tag Sequence:
['B','B','O','B','B','B','O','O','B','B','B','O']

The Internal Revenue Service plans to restructure itself more like a private corporation.
True Tag Sequence:
['B','I','I','I','B','I','I','B','B','B','B','I','I','O']
Predicted Tag Sequence:
['B','I','I','I','I','B','I','B','B','B','B','I','I','B']

Few American officials were willing any longer to defend him .
True Tag Sequence:
['B','I','I','B','I','I','I','I','I','B','O']
Predicted Tag Sequence:
['B','I','I','B','B','B','I','B','I','B','O']

They 're just as confused . ''
True Tag Sequence:
['B','B','B','I','I','O','O']
Predicted Tag Sequence:
['B','B','B','B','B','O','O']

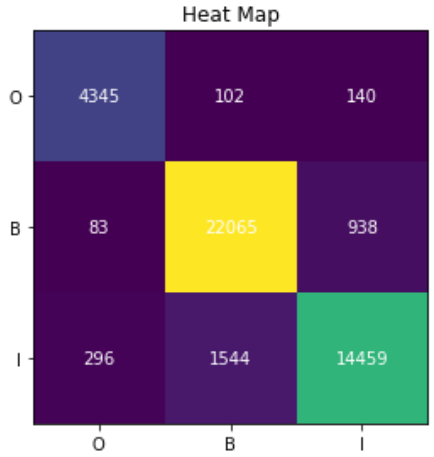
BI-LSTM

Three Methods have been tried:
Each unique element being the word, each unique element being (word, POS tag), each unique element being POS tag

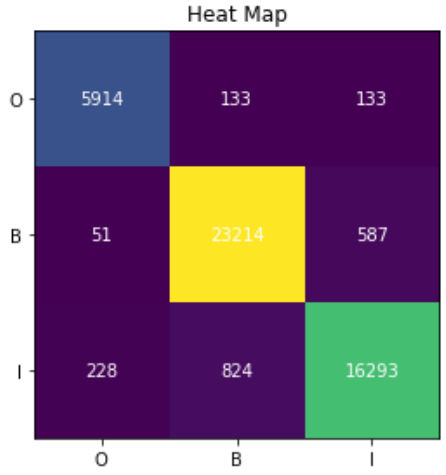
We get the best results if we directly use the POS Tags.

Confusion Matrices

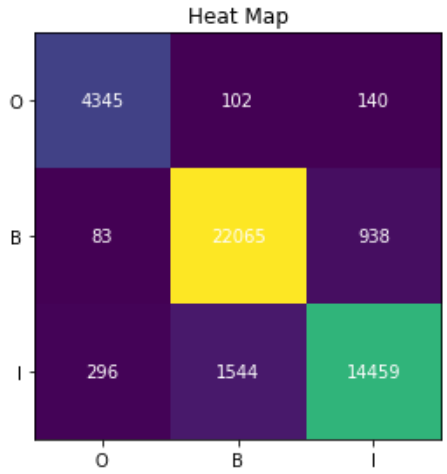
1. Using only the word:



2. Using (word, POS tag)



3. Using only POS tag



Classification Report:

1. Using only the word:

	precision	recall	f1-score	support
O	0.94	0.95	0.95	6180
B	0.95	0.93	0.94	23852
I	0.90	0.92	0.91	17345
accuracy			0.93	47377
macro avg	0.93	0.93	0.93	47377
weighted avg	0.93	0.93	0.93	47377

2. Using (word, tag)

	precision	recall	f1-score	support
O	0.94	0.96	0.95	6180
B	0.93	0.96	0.94	23852
I	0.93	0.88	0.91	17345
accuracy			0.93	47377
macro avg	0.93	0.93	0.93	47377
weighted avg	0.93	0.93	0.93	47377

3. Using only POS tag

	precision	recall	f1-score	support
O	0.95	0.96	0.96	6180
B	0.96	0.97	0.97	23852
I	0.96	0.94	0.95	17345
accuracy			0.96	47377
macro avg	0.96	0.96	0.96	47377
weighted avg	0.96	0.96	0.96	47377

Abalitive Study:

On using POS tags, we get slightly better results with the word ie. an overall accuracy change from 92.5 to 92.83%. Purely using POS tags significantly improves the results, with the overall accuracy going to 95.92%.

After seeing this, we can conclude that in the case of Bi-LSTM, POS tag provides enough information for chunking

Misclassification Examples:

When just using POS tags

Deeply ingrained in both the book review ``Kissing Nature Good-bye'' by Stephen MacDonald -LRB- Leisure & Arts, Sept. 27 -RRB- and the books reviewed is the assumption that global warming is entirely a result of human activity.

Expected:

```
['B', 'I', 'B', 'O', 'B', 'I', 'I', 'O', 'B', 'B', 'B', 'O', 'B', 'B', 'I', 'O',
'B', 'I', 'I', 'O', 'B', 'I', 'O', 'O', 'B', 'I', 'B', 'B', 'B', 'I', 'B', 'B',
'I', 'B', 'B', 'B', 'I', 'B', 'B', 'I', 'O']
```

Tags received:

```
['B', 'I', 'B', 'B', 'I', 'I', 'I', 'O', 'B', 'B', 'I', 'O', 'B', 'B', 'I', 'O',
'B', 'O', 'B', 'O', 'B', 'I', 'O', 'O', 'B', 'I', 'B', 'B', 'B', 'I', 'B', 'B',
'I', 'B', 'B', 'B', 'I', 'B', 'B', 'I', 'O']
```

It has been seen that the errors increase when the model should predict O tags inside the text and the part of speech doesn't correspond to a punctuation. This would imply that previous history from both the left and right of the word gets mixed up when an 'O' tagged word is in proximity.

When using Only the word:

it is probably continuing and may well account for most of, or all of, present-day global warming.

Expected:

```
['O', 'B', 'B', 'I', 'I', 'O', 'B', 'I', 'I', 'B', 'B', 'B', 'O', 'O', 'B', 'B',
'O', 'B', 'I', 'I', 'O']
```

Tags received:

```
['O', 'B', 'B', 'B', 'B', 'I', 'I', 'I', 'I', 'B', 'B', 'B', 'O', 'O', 'B', 'I',
'O', 'B', 'I', 'I', 'O']
```

This gets solved when using POS. Therefore, it is safe to assume that the POS ambiguity has led to the difference in the answer.

CRF Error Analysis (ablative study)

Sentence 1:

Rockwell said the agreement calls for it to supply **200** additional so-called shipsets for the planes.

Actual chunk sequence:

['B', 'B', 'B', 'I', 'B', 'B', 'B', 'B', 'I', 'B', 'I', 'I', 'I', 'B', 'B', 'I', 'O']

Predicted chunk sequence when POS tags excluded:

['B', 'B', 'B', 'I', 'B', 'B', 'B', 'B', 'I', 'I', 'I', 'I', 'I', 'B', 'B', 'I', 'O']

Predicted chunk sequence when POS tags included:

['B', 'B', 'B', 'I', 'B', 'B', 'B', 'B', 'I', 'B', 'I', 'I', 'I', 'B', 'B', 'I', 'O']

Sentence 2:

`Under the existing contract, Rockwell said, it has already delivered 793 of the shipsets to Boeing.`

Actual chunk sequence:

['B', 'B', 'I', 'I', 'O', 'B', 'B', 'O', 'B', 'B', 'I', 'I', 'B', 'B', 'B', 'I', 'B', 'B', 'O']

Predicted chunk sequence when POS tags excluded:

['B', 'B', 'I', 'I', 'O', 'B', 'B', 'O', 'B', 'B', 'I', 'I', 'B', 'B', 'B', 'I', 'B', 'I', 'O']

Predicted chunk sequence when POS tags included:

['B', 'B', 'I', 'I', 'O', 'B', 'B', 'O', 'B', 'B', 'I', 'I', 'B', 'B', 'B', 'I', 'B', 'B', 'O']

Sentence 3:

`To focus on its global consumer-products business, Colgate sold its Kendall health-care business in 1988 .`

Actual chunk sequence:

['B', 'I', 'B', 'B', 'I', 'I', 'I', 'O', 'B', 'B', 'B', 'I', 'I', 'I', 'B', 'B', 'O']

Predicted chunk sequence when POS tags excluded:

['B', 'I', 'B', 'B', 'I', 'I', 'I', 'O', 'B', 'I', 'B', 'I', 'I', 'I', 'B', 'B', 'O']

Predicted chunk sequence when POS tags included:

['B', 'I', 'B', 'B', 'I', 'I', 'I', 'O', 'B', 'B', 'B', 'I', 'I', 'I', 'B', 'B', 'O']

Sentence 4:

By yesterday’s close of trading it was good for a paltry **$43.5** million.

Misclassification errors (predicted correctly by other model namely Bi-LSTM)

Mr. Carlucci, 59 years old, served as defense secretary in the Reagan administration.