Model	P	R	F1	W-F1
PMOA-cite				
SPC + Roberta	0.890	0.845	0.867	0.949
ACL-cite				
SSM (16 sentences) + Roberta	0.807	0.650	0.720	0.939
SSM (16 sentences + section) + Roberta	0.806	0.660	0.726	0.940

Table 1: Performance on Validation Set. SPC: Sentence-Pair Classification, SSM: Sentence Sequence Modeling. P, R, and F-1 are precision, recall, and F-1 scores for the cite class (l_c). W-F1 is the weighted-F1 on entire test dataset.

Layer	Total Parameters
SC	
Roberta Contextual Embedding	125 Million
Feed Forward Network	768
SPC	
Roberta Contextual Embedding	125 Million
Feed Forward Network	768
SSM	
Roberta Contextual Embedding	125 Million
Bi-LSTM	2.7 Million
Feed Forward Network	256

Table 2: Total number of parameters by layers for all three architectures. SC: Sentence Classification, SPC: Sentence-Pair Classification, SSM: Sentence Sequence Modeling.

A Regular expressions

Following are the regular expressions used for extracting citations from from the ACL corpus:

```
author="(?:[\p{Lu}][\p{Lu}\p{Ll}'`-]+)"
etal="(?:et al.?)"
additional="(?:,? (?:(?:and |& )?"+
            author+"|"+etal+"))"

year_num="(?:19|20)[0-9][0-9][a-z]{0,1}"
page_num="(?:, p.? [0-9]+)?"
yp=year_num+page_num
year="(?:[,]{0,1} *"+yp+
      "| *[\[\(](?:[,]{0,1} *"+yp+
      ")+[\)\]])+"

Regex="("+author+additional+"*"+year+")"
```

Below are examples of some types of citations captured by these expressions:

- Brown et al (1993)
- (Langlais et al, 2000)
- (Wu, 1994)
- Wu (1995)
- (Brown et al, 1991; Gale and Church, 1991)
- (Kay and Roscheisen, 1993; Chen, 1993)
- (Brown, 1999a)

B Performance on Validation Set

Table 1 illustrates the performance achieved by the proposed models on the validation sets for PMOA-cite and ACL-cite datasets. Validation numbers are very close to the ones achieved on the test set and the trends are also in-line with the earlier results.

C Total Parameters

Table 2 breaks down the total number of parameters by different Layers for the three proposed architectures.

D Hyper-parameter Selection

We performed manual hyper-parameter selection and picked best parameter values based on the performance on unseen validation set.