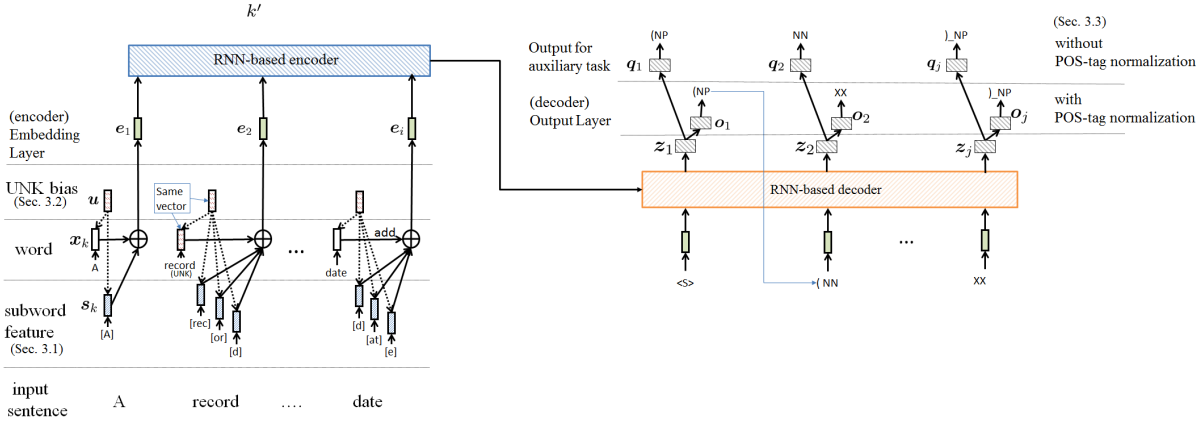
Supplementary Material of "An Empirical Study of Building a Strong Baseline for Constituency Parsing"

A Detailed explanation of the model used in our experiments

$$e_k = E(x_i + u) + \sum_{k'} F(s_{k'} + u) \qquad o_j = W^{(o)} z_j \qquad q_j = W^{(q)} z_j$$



This figure (Figure A) shows the sketch of how we incorporate several generic techniques explained in Sec.3: subword features (§3.1), unknown token embeddings (§3.2), and jointly estimating POS-tags (with bracketing) as an auxiliary task of multitask learning (§3.3).

B Actual Evaluation Results

The following subsections from B.1 through B.6 show the actual outputs of the evalb evaluation script in our experiments.

Note that all the results reported in this paper get zero for both "Number of Error sentence" and "Number of Skip sentence". This is essentially very important to confirm since there is a known issue in the evalb evaluation script that it simply discards malformed (error) outputs from the evaluation. As a result, the system with many malformed outputs gets better performance since the malformed (error) outputs from the evaluation.. the parser that generates the better scores

B.1 Actual evaluation output of `(e)`: `(d)` + Pos in Table 4

```
$ ./EVALB/evalb -p EVALB/COLLINS.prm   data/sec23.gold  ${FILE} | tail -n31
============================================================================
               90.22  91.02  39948 44276 43889   1872  49892 49892   100.00
=== Summary ===

-- All --
Number of sentence        =   2416
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2416
Bracketing Recall         =  90.22
Bracketing Precision      =  91.02
Bracketing FMeasure       =  90.62
Complete match            =  41.39
Average crossing          =   0.77
No crossing               =  69.50
2 or less crossing        =  89.82
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   2245
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2245
Bracketing Recall         =  90.83
Bracketing Precision      =  91.63
```

```
Bracketing FMeasure       =  91.23
Complete match            =  43.83
Average crossing          =   0.64
No crossing               =  71.98
2 or less crossing        =  91.67
Tagging accuracy          = 100.00
```

B.2 Actual evaluation output of `(j)`: `(i)` + Pos in Table 4

```
$ ./EVALB/evalb -p EVALB/COLLINS.prm   data/sec23.gold  ${FILE} | tail -n31
============================================================================
                91.01  91.72  40294 44276 43930   1748  49892 49892   100.00
=== Summary ===

-- All --
Number of sentence        =   2416
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2416
Bracketing Recall         =  91.01
Bracketing Precision      =  91.72
Bracketing FMeasure       =  91.36
Complete match            =  43.50
Average crossing          =   0.72
No crossing               =  71.23
2 or less crossing        =  90.65
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   2245
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2245
Bracketing Recall         =  91.65
Bracketing Precision      =  92.33
Bracketing FMeasure       =  91.99
Complete match            =  46.10
Average crossing          =   0.58
No crossing               =  73.85
2 or less crossing        =  92.52
Tagging accuracy          = 100.00
```

B.3 Actual evaluation output of `(k)`: `(e)` + ensemble $A = 8$ shown in Table 5

```
$ ./EVALB/evalb -p EVALB/COLLINS.prm   data/sec23.gold  ${FILE} | tail -n31
============================================================================
                91.81  92.55  40648 44276 43920   1527  49892 49892   100.00
=== Summary ===

-- All --
Number of sentence        =   2416
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2416
Bracketing Recall         =  91.81
Bracketing Precision      =  92.55
Bracketing FMeasure       =  92.18
Complete match            =  45.90
Average crossing          =   0.63
No crossing               =  73.43
2 or less crossing        =  91.64
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   2245
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2245
Bracketing Recall         =  92.21
Bracketing Precision      =  92.95
```

```
Bracketing FMeasure       =  92.58
Complete match            =  48.33
Average crossing          =   0.53
No crossing               =  75.68
2 or less crossing        =  93.05
Tagging accuracy          = 100.00
```

B.4 Actual evaluation output of `(l) (k)` + LM-rerank $C = 80$ shown in Table 5

```
$ ./EVALB/evalb -p EVALB/COLLINS.prm   data/sec23.gold  ${FILE} | tail -n31
============================================================================
               93.72  94.57  41496 44276 43880   1042  49892 49892   100.00
=== Summary ===

-- All --
Number of sentence        =   2416
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2416
Bracketing Recall         =  93.72
Bracketing Precision      =  94.57
Bracketing FMeasure       =  94.14
Complete match            =  52.69
Average crossing          =   0.43
No crossing               =  80.50
2 or less crossing        =  94.37
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   2245
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2245
Bracketing Recall         =  94.12
Bracketing Precision      =  94.93
Bracketing FMeasure       =  94.53
Complete match            =  55.32
Average crossing          =   0.35
No crossing               =  82.45
2 or less crossing        =  95.81
Tagging accuracy          = 100.00
```

B.5 Actual evaluation output of `(m)`: `(j)` + ensemble $A = 8$ shown in Table 5

```
$ ./EVALB/evalb -p EVALB/COLLINS.prm   data/sec23.gold  ${FILE} | tail -n31
============================================================================
               92.35  93.13  40890 44276 43907   1392  49892 49892   100.00
=== Summary ===

-- All --
Number of sentence        =   2416
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2416
Bracketing Recall         =  92.35
Bracketing Precision      =  93.13
Bracketing FMeasure       =  92.74
Complete match            =  47.27
Average crossing          =   0.58
No crossing               =  75.17
2 or less crossing        =  92.26
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   2245
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2245
Bracketing Recall         =  92.77
```

```
Bracketing Precision      =  93.55
Bracketing FMeasure       =  93.16
Complete match            =  49.53
Average crossing          =   0.47
No crossing               =  77.37
2 or less crossing        =  93.99
Tagging accuracy          = 100.00
```

B.6 Actual evaluation output of `(n)`: `(m)` + LM-rerank $C = 80$ shown in Table 5

```
$ ./EVALB/evalb -p EVALB/COLLINS.prm   data/sec23.gold  ${FILE} | tail -n31
============================================================================
                93.91  94.72  41580 44276 43896   1014  49892 49892   100.00
=== Summary ===

-- All --
Number of sentence        =   2416
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2416
Bracketing Recall         =  93.91
Bracketing Precision      =  94.72
Bracketing FMeasure       =  94.32
Complete match            =  52.81
Average crossing          =   0.42
No crossing               =  80.92
2 or less crossing        =  94.66
Tagging accuracy          = 100.00

-- len<=40 --
Number of sentence        =   2245
Number of Error sentence  =      0
Number of Skip  sentence  =      0
Number of Valid sentence  =   2245
Bracketing Recall         =  94.25
Bracketing Precision      =  95.03
Bracketing FMeasure       =  94.64
Complete match            =  55.41
Average crossing          =   0.35
No crossing               =  82.72
2 or less crossing        =  95.95
Tagging accuracy          = 100.00

============================================================================
```


Acknowledgement

We thank three anonymous reviewers for their insightful and valuable comments.