Supplementary Material for “An Empirical Investigation of Structured Output Modeling for Graph-based Neural Dependency Parsing”

A Hyper-Parameter Settings

Table 1 summarizes the hyper-parameters of our models. For the inputs, we concatenate word, Part-of-Speech (POS) embeddings and character-level representations from a Char-CNN with a window size of five. Three layers of BiLSTM are utilized to obtain contextual representations. Before feeding into the biaffine scorer, the representations are further transformed using feed-forward layers for arc-scoring and label-scoring separately. Arc scores and label scores are combined for the final output scores. The architecture is basically mostly previous work (Dozat and Manning, 2017), and the main focus of our exploration is the final output layer without any explicit neural parameters.

Layer	Hyper-Parameter	Value
Word	dimension	300
POS	dimension	50
Char-CNN	dimension	30
Encoder	type	BiLSTM
	encoder layer	3
	encoder size	512
MLP	arc MLP size	512
	label MLP size	128
Training	Dropout	0.33
	optimizer	Adam
	learning rate	0.001
	batch size	32

Table 1: Hyper-parameters in our experiments.

B Results of Unlabeled Scores

Method	Single	Local	Global-NonProj		Global-Proj	
	Prob	Prob	Prob	Hinge	Prob	Hinge
PTB	95.25/55.68	95.52/57.15	95.59/58.39†	95.64†/58.50†	95.58/59.19†	**95.68**†/**59.75**†
CTB	88.27/35.27	89.48/37.59	89.52/38.27	89.29/37.42	**89.65**†/**39.41**†	89.48/38.74†
bg-btb	94.00/59.62	93.86/59.59	94.12†/60.87†	**94.27**†/61.95†	93.92/**62.13**†	94.05†/62.07†
ca-ancora	93.39/35.95	93.51/36.11	93.59†/36.96†	**93.71**†/**37.09**†	93.19/36.87	93.43/36.80
cs-pdt	93.84/57.37	94.20/59.11	**94.28**†/**59.86**†	94.22/59.63†	93.70/57.01	93.84/56.88
de-gsd	**88.35**†/38.83	88.03/37.97	88.10/38.93†	88.23†/38.25	87.88/**39.88**†	88.28†/38.59
en-ewt	89.85/61.88	90.26/62.48	90.34/63.60†	**90.46**†/63.22†	90.34/**64.92**†	90.32/63.91†
es-ancora	92.78/36.47	92.92/36.37	92.96/37.28†	**93.05**†/36.76	92.62/**37.42**†	92.92/37.13
fr-gsd	91.17/33.09	91.09/32.77	91.29/34.38	91.32†/33.65	**91.46**†/**36.46**†	91.43†/34.94†
it-isdt	93.63/53.80	93.83/53.94	93.84/54.70	93.84/54.56	94.08†/**58.02**†	**94.11**†/56.71†
nl-alpino	91.22/43.62	91.56/43.68	91.71/45.13	**91.81**†/**45.92**†	91.10/41.55	91.31/43.01
no-bokmaal	94.31/60.44	94.27/60.48	**94.35/61.30**†	94.26/60.75	94.17/60.79	94.16/60.65
ro-rrt	90.82†/31.18	90.38/29.95	90.49/31.92†	90.70†/30.50	90.54/32.42†	**90.86**†/**32.78**†
ru-syntagrus	94.14/56.37	94.57/58.22	94.62/58.55	**94.62/58.74**†	94.33/58.04	94.50/57.82
Average	92.21/47.11	92.39/47.53	92.49†/48.58†	**92.53**†/48.35†	92.33/**48.86**†	92.46/48.56†

Table 2: Unlabeled results (**UAS/UCM**) on the test sets (averaged over three runs). ‘†’ means that the result of the model is statistically significantly better (by permutation test, $p < 0.05$) than the Local-Prob model. The patterns are similar to the ones listed in the main content.

C Details of Data

Our experiments are performed on English Penn Treebank (PTB), Penn Chinese Treebank (CTB) and 12 selected treebanks from Universal Dependencies (v2.3) (Nivre et al., 2018). We follow standard data preparing conventions: For PTB, we follow the dataset splitting convention: Sections 2-21 for training, Section 22 for validation and Section 23 for testing. Dependency trees are obtained using the converter in Stanford Parser version 3.3.0. The POS tags were predicted using the Stanford POS tagger (Toutanova et al., 2003) with 10-fold jackknifing on the training data. For CTB, we follow the splitting of (Zhang and Clark, 2008) and the dependencies are converted using the Penn2Malt converter. Following previous work, gold segmentation and POS tags are used. For UD, we select 12 relatively large treebanks of UD version 2.3 (Nivre et al., 2018), and also use gold POS tags.

For evaluation, due to space limitation, we only report LAS (Labeled Attachment Score) and LCM (Labeled Complete Match) in the main content. We also include the unlabeled scores UAS (Unlabeled Attachment Score) and UCM (Unlabeled Complete Match) in the supplementary material. The evaluations on PTB and CTB exclude punctuations (tokens whose gold POS tag is one of {“ ” : , .} for PTB or “PU” for CTB), while on UD we include all tokens.

The details of the selected treebanks are listed in Table 3.

Treebank		#Sent	#Token	Proj-Sent(%)	Proj-Token(%)
PTB	train	39832	950028	99.90	99.99
	dev	1700	40117	99.82	99.98
	test	2416	56684	99.96	100.00
CTB	train	16091	437990	100.00	100.00
	dev	803	20454	100.00	100.00
	test	1910	50315	100.00	100.00
bg-btb	train	8907	124336	96.83	99.22
	dev	1115	16089	97.49	99.36
	test	1116	15724	97.13	99.31
ca-ancora	train	13123	417587	89.90	98.79
	dev	1709	56482	89.12	98.78
	test	1846	57902	89.06	98.65
cs-pdt	train	68495	1173282	88.22	97.19
	dev	9270	159284	87.46	97.04
	test	10148	173918	88.03	97.18
de-gsd	train	13814	263804	90.67	97.94
	dev	799	12486	93.24	98.48
	test	977	16498	90.69	97.73
en-ewt	train	12543	204585	94.67	99.01
	dev	2002	25148	97.05	99.32
	test	2077	25096	96.53	99.12
es-ancora	train	14305	444617	90.49	99.00
	dev	1654	52336	90.02	98.92
	test	1721	52617	90.35	98.94
fr-gsd	train	14450	354699	91.94	99.06
	dev	1476	35720	93.02	99.22
	test	416	10021	95.19	99.47
it-isdt	train	13121	276019	98.01	99.71
	dev	564	11908	96.63	99.55
	test	482	10417	96.68	99.47
nl-alpino	train	12269	186046	85.57	95.64
	dev	718	11541	90.81	97.69
	test	596	11046	85.74	96.59
no-bokmaal	train	15696	243887	92.15	98.12
	dev	2410	36369	92.66	98.10
	test	1939	29966	92.37	98.04
ro-rrt	train	8043	185113	88.61	98.32
	dev	752	17074	88.56	98.28
	test	729	16324	90.26	98.55
ru-syntagrus	train	48814	870474	92.00	98.31
	dev	6584	118487	92.13	98.38
	test	6491	117329	92.05	98.35

Table 3: Statistics of the Treebanks. “Proj-Sent” and “Proj-Token” denote projective rate of the sentences and tokens, perspectively.

References

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *ICLR*.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2008. A tale of two parsers: investigating and combining graph-based and transition-based dependency parsing using beam-search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 562–571. Association for Computational Linguistics.