Second-Order Neural Dependency Parsing with Message Passing and End-to-End Training

Xinyu Wang and Kewei Tu

ShanghaiTech University

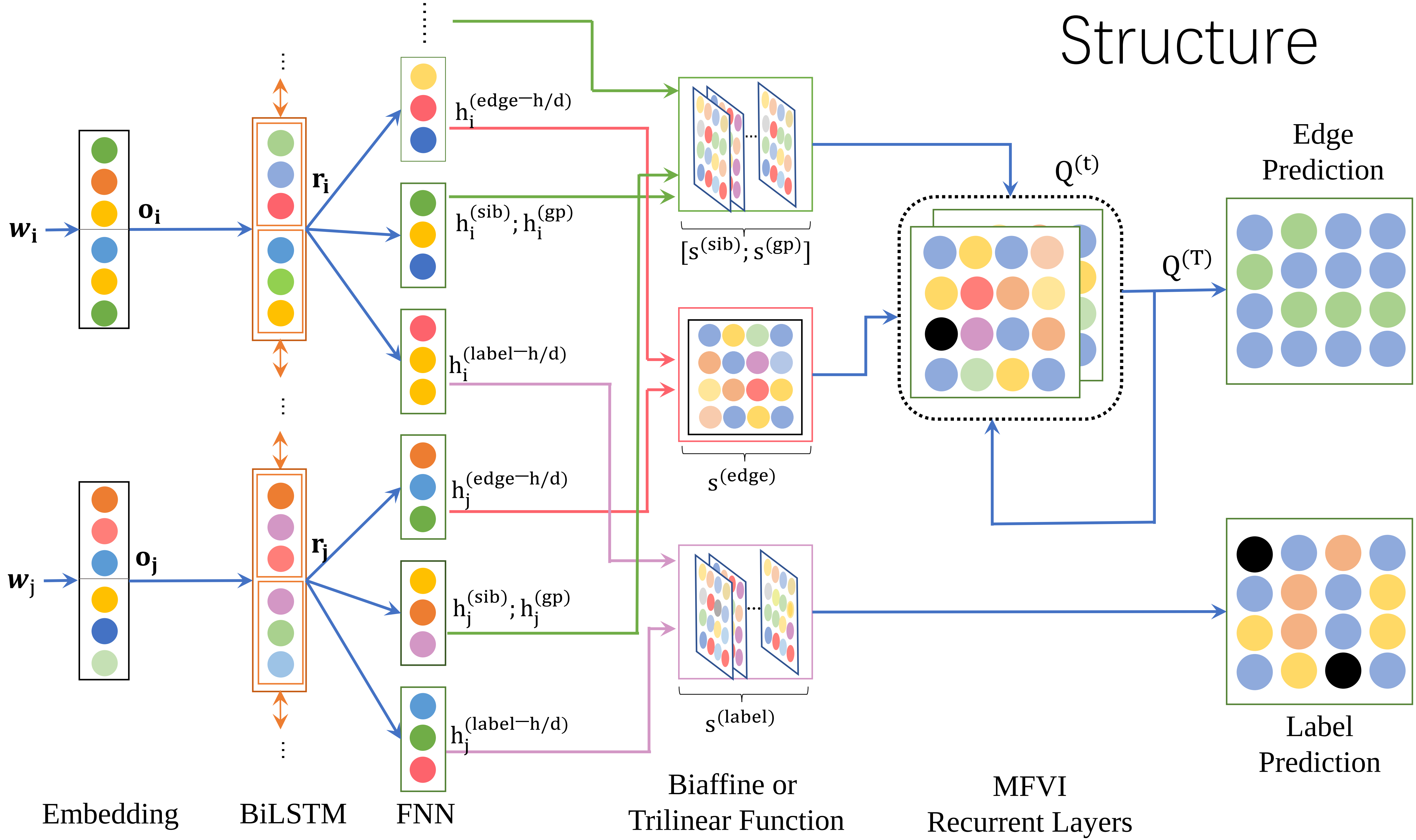
上海科技大学

ShanghaiTech University

Motivation and Contributions

- Higher-order approaches have achieved state-of-the-art performance
- Our work:
  - Apply second-order semantic parser (Wang et al., 2019) to syntactic dependency parsing.
- Our observation:
  - Higher-order decoding is effective even with contextual word embeddings.
  - Parsers without head-selection constraint can match the accuracy of parsers with the head-selection constraint and can even outperform the latter when using BERT embedding

Xinyu Wang, Jingxian Huang, and Kewei Tu. Second-order semantic dependency parsing with end-to-end neural networks. In ACL 2019


Structure
w_i
o_i
r_i
w_j
o_j
r_j
$h_i^{(edge-h/d)}$
$h_i^{(sib)}; h_i^{(gp)}$
$h_i^{(label-h/d)}$
$h_j^{(edge-h/d)}$
$h_j^{(sib)}; h_j^{(gp)}$
$h_j^{(label-h/d)}$
$[s^{(sib)}; s^{(gp)}]$
$s^{(edge)}$
$s^{(label)}$
$Q^{(t)}$
$Q^{(T)}$
Edge Prediction
Label Prediction
Embedding
BiLSTM
FNN
Biaffine or Trilinear Function
MFVI Recurrent Layers

Approach

Binary Classification (Single):

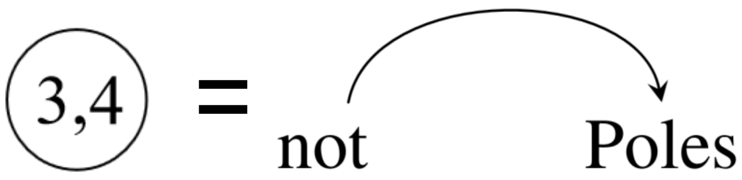
$$\mathcal{M}_{ij}^{(t-1)} = \sum_{k \neq i,j} Q_{ik}^{(t-1)}(1) s_{ij,ik}^{(sib)} + Q_{jk}^{(t-1)}(1) s_{ij,jk}^{(gp)} + Q_{ki}^{(t-1)}(1) s_{ki,ij}^{(gp)}$$

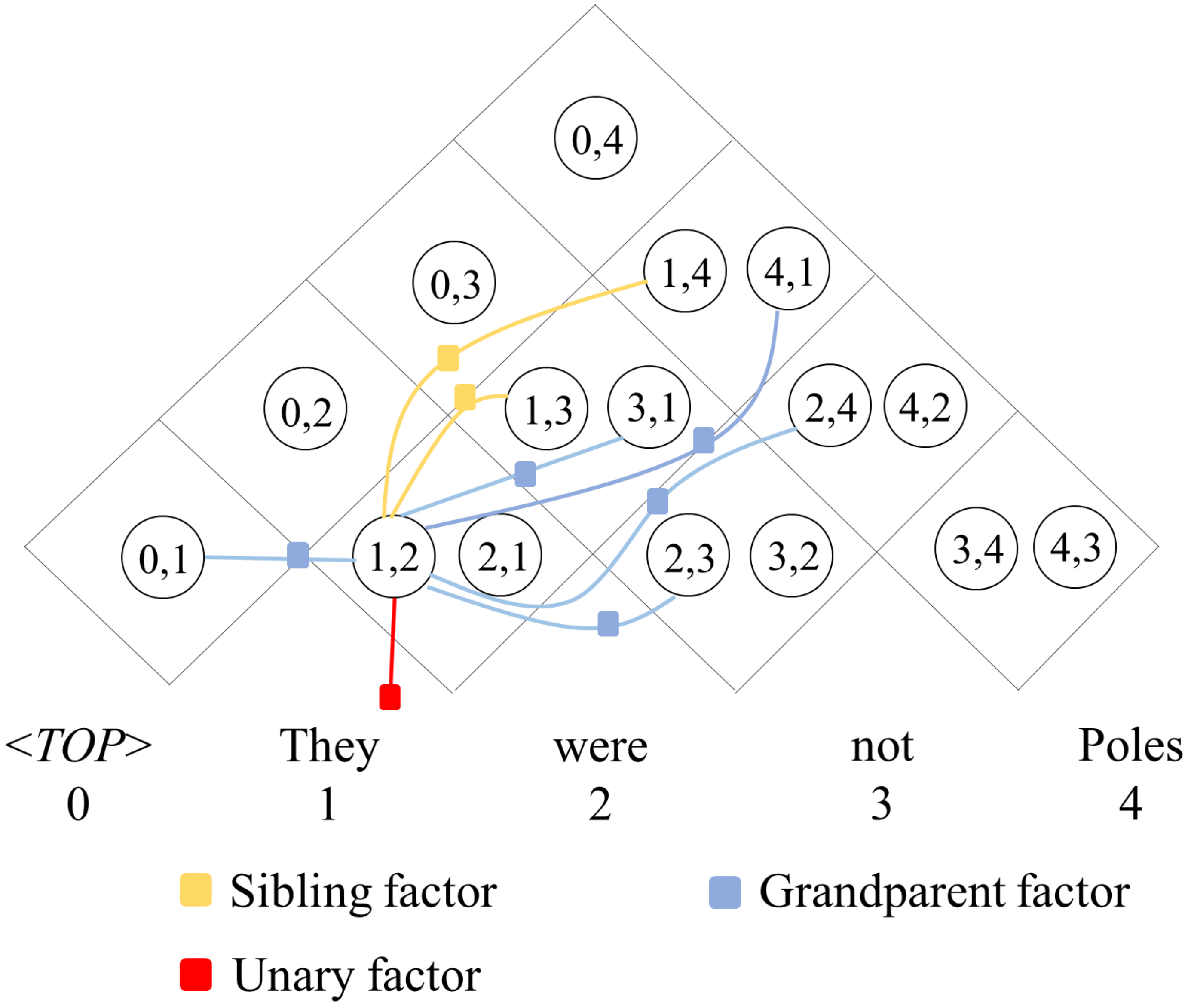
$$Q_{ij}^{(t)}(0) \propto 1$$

$$Q_{ij}^{(t)}(1) \propto \exp\{s_{ij}^{(\text{edge})} + \mathcal{M}_{ij}^{(t-1)}\}$$

Conditional Random Field

Nodes: Edges between two words





Approach

Binary Classification (Single):

$$\mathcal{M}_{ij}^{(t-1)} = \sum_{k \neq i,j} Q_{ik}^{(t-1)}(1) s_{ij,ik}^{(sib)} + Q_{jk}^{(t-1)}(1) s_{ij,jk}^{(gp)} + Q_{ki}^{(t-1)}(1) s_{ki,ij}^{(gp)}$$

$$Q_{ij}^{(t)}(0) \propto 1$$

$$Q_{ij}^{(t)}(1) \propto \exp\{s_{ij}^{(\text{edge})} + \mathcal{M}_{ij}^{(t-1)}\}$$

Head-selection (Local):

$$\mathcal{M}_{j}^{(t-1)}(i) = \sum_{k \neq i,j} Q_{k}^{(t-1)}(i) s_{ij,ik}^{(sib)} + Q_{k}^{(t-1)}(j) s_{ij,jk}^{(gp)} + Q_{i}^{(t-1)}(k) s_{ki,ij}^{(gp)}$$

$$Q_{j}^{(t)}(i) = \frac{\exp\{s_{ij}^{(\text{edge})} + \mathcal{M}_{j}^{(t-1)}(i)\}}{\sum_{k=0}^{n} \exp\{s_{kj}^{(\text{edge})} + \mathcal{M}_{j}^{(t-1)}(k)\}}$$

Results

	PTB		CTB	
	UAS	LAS	UAS	LAS
Dozat and Manning (2017)	95.74	94.08	89.30	88.23
Ma et al. (2018)[♠]	95.87	94.19	90.59	89.29
F&G (2019)[♠]	96.04	94.43	-	-
GNN	95.87	94.15	90.78	89.50
Single1O	95.75	94.04	90.53	89.28
Local1O	95.83	94.23	90.59	89.28
Single2O	95.86	94.19	90.75	89.55
Local2O	95.98	94.34	**90.81**	**89.57**
Ji et al. (2019)[†]	95.97	94.31	-	-
Zhang et al. (2020)[†‡]	**96.14**	**94.49**	-	-
Local2O[†‡]	96.12	94.47	-	-
+BERT				
Zhou and Zhao (2019)[♣]	97.20	95.72		
Clark et al. (2018)[◇]	96.60	95.00	-	-
Single1O	96.82	95.20	92.73	91.64
Local1O	96.86	95.32	92.47	91.30
Single2O	96.86	95.31	**92.78**	**91.69**
Local2O	**96.91**	**95.34**	92.55	91.38

Results

	PTB	CTB	bg	ca	cs	de	en	es	fr	it	nl	no	ro	ru	Avg.
GNN	94.15	89.50†	90.33	92.39	90.95	79.73	88.43	91.56	87.23	92.44	88.57	89.38	85.26	91.20	89.37
Single1O	94.04	89.28	90.05	92.72†	92.07	81.73	89.55	92.10	88.27	92.64	89.57	91.81	85.39	92.60	90.13
Local1O	94.23	89.28	90.30	92.56	**92.15**	81.42	89.43	91.99	88.26	92.49	89.76	**91.91**	85.27	**92.72**	90.13
Single2O	94.19	89.55†	90.24	92.82†	92.13	**81.99**†	89.64†	**92.17**†	**88.69**	**92.83**†	89.97†	91.90	85.53†	92.58	90.30†
Local2O	**94.34**†‡	**89.57**†	**90.53**†	**92.83**†	92.12	81.73	**89.72**†	92.07	88.53	92.78	**90.19**†	91.88	**85.88**†‡	92.67	**90.35**†
+BERT															
Single1O	95.20	91.64†	90.87	93.55†	92.01	81.95†	90.44†	92.56†	**89.35**	93.44†	90.89	91.78	86.13†	92.51	90.88†
Local1O	95.32	91.30	91.03	93.17	91.93	81.66	90.09	92.32	89.26	93.05	90.93	91.62	85.67	92.51	90.70
Single2O	95.31	**91.69**†‡	**91.30**†	**93.60**†‡	**92.09**†	**82.00**†‡	**90.75**†‡	**92.62**†‡	89.32	**93.66**†	**91.21**	**91.74**	**86.40**†	92.61	**91.02**†‡
Local2O	**95.34**	91.38	91.13	93.34†	92.07†	81.67	90.43†	92.45†	89.26	93.50†	90.99	91.66	86.09†	**92.66**	90.86†

† means that the model is statistically significantly better than the Local1O model with a significance level of $p<0.05$
‡ represents winner of the significant test between the Single2O and Local2O models

- Our second-order approaches outperform GNN and the first-order approaches both with and without BERT embeddings
- Without BERT, Local approaches slightly outperforms Single approaches, although the difference between the two is quite small
- When BERT is used, Single approaches clearly outperforms Local approaches
- The relative strength of Local and Single approaches varies over treebanks, suggesting varying importance of the head-selection constraint

Speed Comparison

(Sentences/Second)

System	Train	Test	Time Complexity
GNN	392	464	$O(n^2 d)$
Zhang et al. (2020)	200	400	$O(n^3)$
Single1O	616	1123	$O(n^2)$
Local1O	**625**	**1150**	$O(n^2)$
Single2O	481	966	$O(n^3)$
Local2O	486	1006	$O(n^3)$

Conclusion

- Second-order graph-based dependency parsing based on message passing and end-to-end neural networks
- Design a new approach that incorporates the head-selection structured constraint
- Show the effectiveness of second-order parsers against first-order parsers even with contextual embeddings
- Competitive accuracy with recent SOTA second-order parsers and significantly faster speed
- The limited usefulness of the head-selection constraint