University of Electronic Science and Technology of China
HUAZHONG UNIVERSITY SCIENCE AND TECHNOLOGY

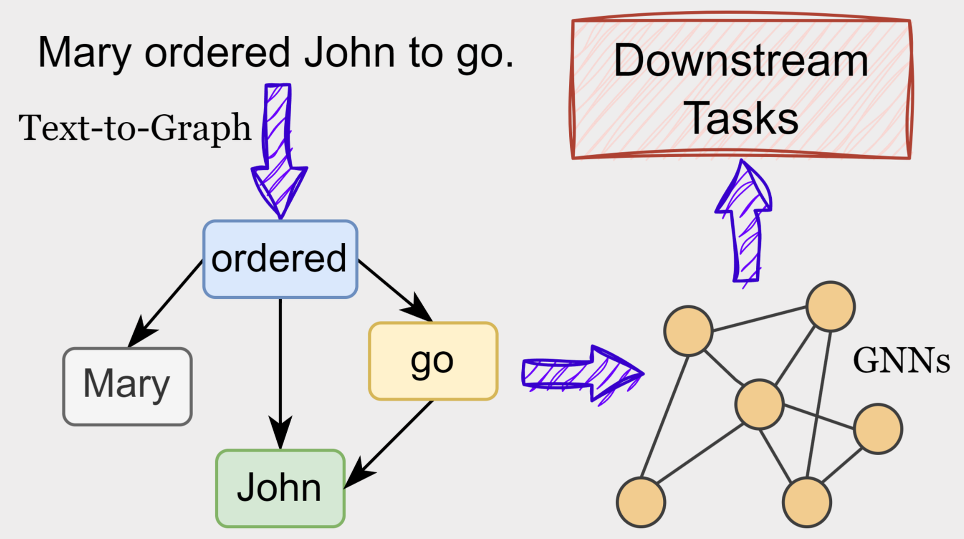
ICASSP 2024 KOREA

MLPs Compass: What is learned when MLPs are combined with PLMs?

Li Zhou, Wenyu Chen, Yong Cao, Dingyi Zeng, Wanlong Liu, Hong Qu

Motivation

Traditional method



Observation

Performance improves when applying MLPs **without structural bias** on pre-trained language models for the relation extraction task

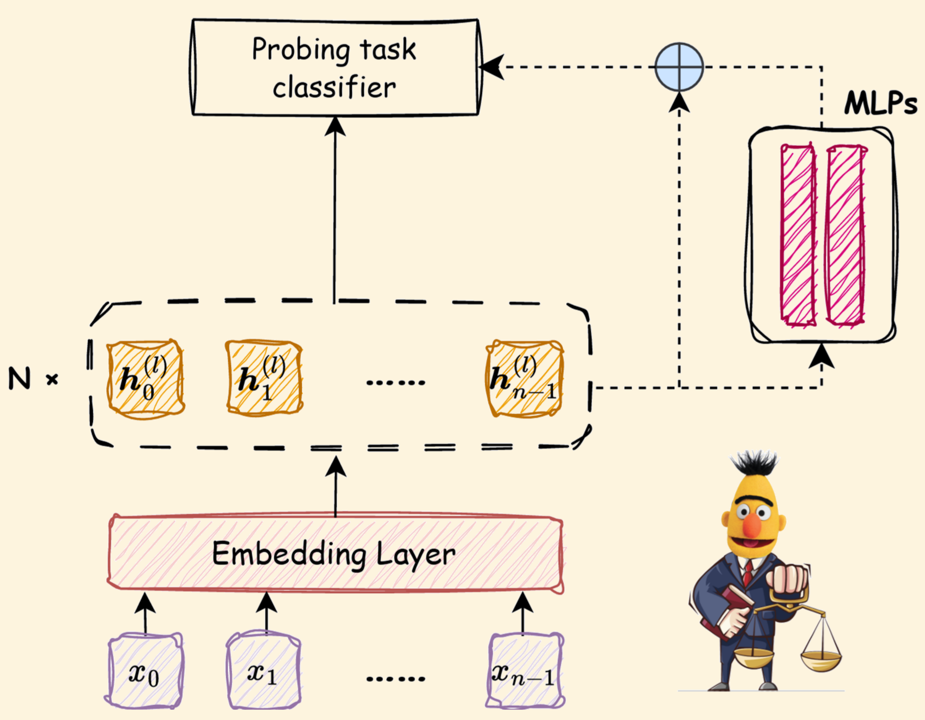
	ReTACRED	SemEval
BERT	87.66±0.18	91.07±0.26
BERT+MLPs	88.05±0.21	91.31±0.23

Research Question

1) **What can be learned** when basic MLPs are integrated with the transformer structure in PLMs?
2) **Does layer sensitivity exist** in the performance changes when combining MLPs and PLM?
3) In the enhancement of PLMs with MLPs, **which aspect of linguistic information** understanding is MLPs particularly skilled at improving?

Method

Probing Framework



All parameters inside the dashed line and the embedding layer are fixed.

Probing Tasks

Two surface tasks

Three syntactic tasks

Five semantic tasks

Experiment

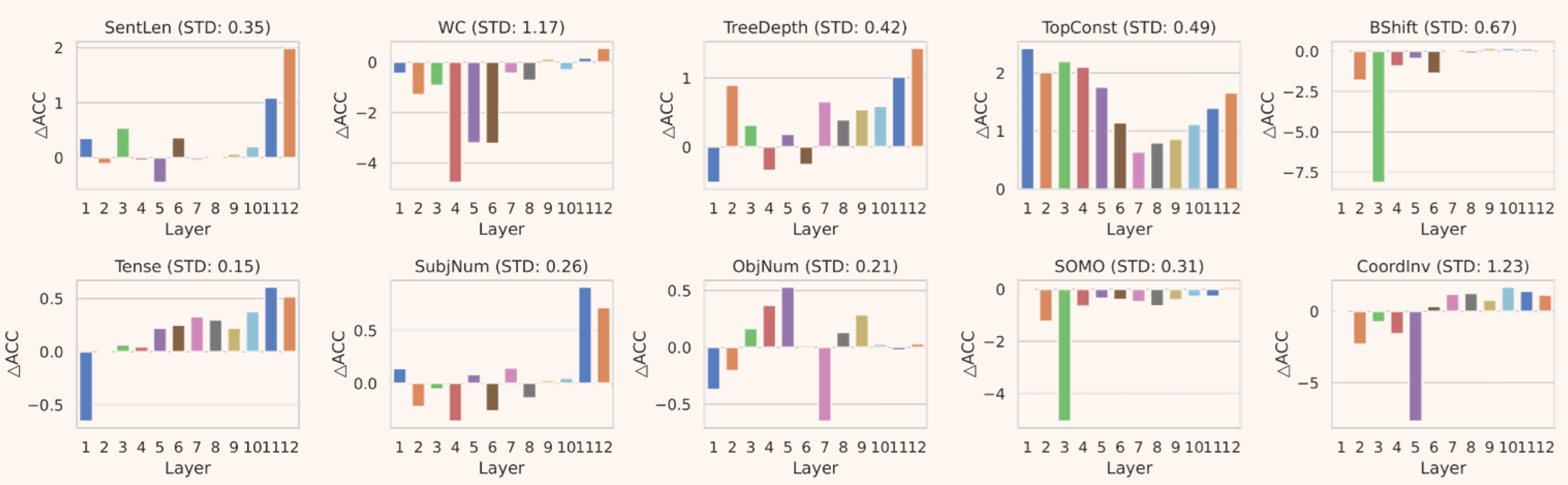
RQ1: Layer-wise Results

Layers	Surface				Syntactic					
	SentLen (6)		WC (1000)		TreeDepth (7)		TopConst (20)		BShift (2)	
	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w
1	85.83±0.95	**86.19±1.17**	0.56±0.05	0.12±0.04	31.60±0.58	31.09±1.17	46.12±0.28	**48.54±0.16**	50.00±0.00	50.01±0.01
2	91.60±0.40	91.49±1.35	2.35±0.10	1.06±0.08	34.68±0.59	**35.58±0.29**	58.19±0.41	**60.2±0.43**	51.81±1.05	50.00±0.00
3	92.31±0.48	**92.85±0.56**	1.50±0.17	0.58±0.05	33.98±0.37	**34.3±0.38**	56.77±0.18	**58.97±0.65**	58.13±1.78	50.00±0.00
4	89.70±0.79	89.66±0.58	19.83±0.71	15.05±0.83	33.08±0.45	32.74±1.60	54.50±0.40	**56.60±0.51**	69.74±1.47	68.83±2.12
5	85.00±0.72	84.55±0.78	19.47±0.62	16.26±0.81	33.90±0.97	**34.08±0.76**	73.93±0.11	**75.69±0.49**	78.44±0.32	77.99±0.40
6	81.10±0.81	**81.46±0.49**	13.79±0.47	10.57±0.74	35.22±0.38	34.97±1.36	78.86±0.13	**80.0±0.50**	80.68±0.14	79.33±1.11
7	78.52±0.86	78.47±0.66	10.33±0.30	9.90±0.33	34.98±0.53	**35.64±0.56**	80.32±0.15	**80.96±0.10**	81.25±0.14	**81.33±0.17**
8	76.99±1.06	**77.01±1.17**	7.99±0.15	7.27±0.19	34.15±0.44	**34.54±0.22**	79.55±0.20	**80.35±0.34**	81.98±0.25	81.86±0.29
9	74.15±0.45	**74.21±0.96**	9.14±0.08	**9.27±0.20**	34.06±0.36	**34.60±0.34**	79.52±0.24	**80.38±0.32**	85.51±0.19	**85.70±0.13**
10	72.82±0.21	**73.01±0.88**	9.41±0.16	9.11±0.36	33.72±0.66	**34.31±0.33**	78.76±0.23	**79.87±0.26**	85.72±0.18	**85.90±0.09**
11	68.88±0.32	**69.96±0.89**	10.59±0.28	**10.75±0.28**	32.75±0.32	**33.76±0.77**	77.02±0.15	**78.42±0.28**	85.86±0.15	**85.98±0.19**
12	64.35±0.26	**66.34±0.89**	14.26±0.24	**14.82±0.54**	31.39±0.39	**32.82±0.46**	72.86±0.16	**74.52±0.13**	86.13±0.08	**86.20±0.30**

Layers	Semantic									
	Tense (2)		SubjNum (2)		ObjNum (2)		SOMO (2)		CoordInv (2)	
	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w
1	78.58±0.25	77.92±0.47	73.39±0.41	**73.53±0.18**	71.08±0.46	70.70±0.75	49.98±0.13	49.97±0.13	50.00±0.00	50.00±0.00
2	84.34±0.27	84.33±0.54	79.02±0.20	78.80±0.23	77.31±0.67	77.11±1.18	51.20±1.08	49.97±0.13	52.31±1.21	50.00±0.00
3	85.45±0.30	**85.51±0.37**	79.44±0.13	79.38±0.20	76.27±1.43	**76.44±0.76**	55.04±0.49	49.97±0.13	50.74±0.95	50.00±0.00
4	86.33±0.34	**86.37±0.49**	79.51±0.23	79.15±0.47	77.73±0.90	**78.10±0.09**	57.88±0.14	57.23±0.35	51.59±0.94	50.00±0.00
5	88.63±0.16	**88.85±0.29**	83.40±0.43	**83.48±0.40**	78.48±0.60	**79.01±0.27**	59.33±0.30	58.98±0.48	57.72±1.15	50.01±0.01
6	88.60±0.28	**88.85±0.27**	86.34±0.24	86.08±0.91	79.12±0.62	**79.13±0.50**	59.68±0.12	59.29±0.28	63.73±1.14	**64.07±0.51**
7	88.86±0.18	**89.19±0.25**	85.76±0.29	**85.91±0.47**	79.73±0.48	79.08±0.19	60.42±0.37	59.94±0.49	69.66±1.05	**70.86±0.95**
8	89.16±0.14	**89.46±0.29**	85.96±0.32	85.82±0.60	79.02±0.26	**79.15±0.33**	60.32±0.42	59.68±0.61	71.14±0.86	**72.41±0.57**
9	89.21±0.08	**89.43±0.26**	86.66±0.11	**86.69±0.23**	79.21±0.40	**79.50±0.12**	62.37±0.14	61.96±0.31	73.74±0.82	**74.53±0.77**
10	89.10±0.08	**89.47±0.21**	85.98±0.26	**86.03±0.14**	78.14±0.26	**78.17±0.38**	62.70±0.19	62.41±0.34	73.82±1.17	**75.52±0.86**
11	88.86±0.31	**89.46±0.20**	83.56±0.50	**84.47±0.25**	77.09±0.23	77.07±0.41	63.55±0.15	63.28±0.30	73.27±0.53	**74.68±0.65**
12	88.87±0.27	**89.39±0.11**	82.26±0.18	**82.97±0.44**	77.88±0.22	**77.91±0.31**	64.00±0.21	**64.09±0.20**	71.25±0.69	**72.38±0.52**

In most layers of the probing experiments, combining MLPs with PLM can improve the performance of the probing tasks at three different levels.

RQ2: Layer Sensitivity



The ability of MLPs to capture additional language information varies across BERT's middle and lower-level layers, while consistently proving beneficial in its higher layers.

RQ3: Linguistic Information Comparison

	Surface	Syntactic	Semantic
NMI (w/o)	0.60	0.14	0.07
NMI (w)	0.66	0.57	0.49
ΔNMI	0.06 (↑)	0.43 (↑)	0.42 (↑)

Clustering performance with Normalized Mutual Information (NMI).

MLPs are better at capturing both syntactic and semantic information compared to surface one.

MyWebsite

PAPER LINK

Conclusion

1. Our extensive experiments, encompassing 10 probing tasks spanning 3 linguistic levels, demonstrate the superior performance of our proposed framework.
2. MLPs can boost PLMs in capturing additional surface, syntactic, and semantic information, with a stronger capacity for enhancing the latter two.
3. When leveraging high-layer representations from PLMs, MLPs exhibit a greater ability to acquire additional information.
4. Our work provides interpretable and valuable insights into crafting variations of PLMs utilizing MLPs for tasks that emphasize diverse linguistic structures.