Pre-training Language Model as a Multi-perspective Course Learner

Beiduo Chen[§‡], Shaohan Huang[‡†], Zihan Zhang[‡], Wu Guo[§],
Zhenhua Ling[§], Haizhen Huang[‡], Furu Wei[‡], Weiwei Deng[‡] and Qi Zhang[‡]

[§] NERC-SLIP, University of Science and Technology of China, Hefei, China [‡] Microsoft Corporation, Beijing, China

Introduction

- **Task** Language model pre-training based on ELECTRA-style framework.
- **Issue** Generator (G) with only MLM leads to biased learning and label imbalance for discriminator (D); no explicit feedback loop from D to G results in the chasm between these two components.
- **Solution** Multi-perspective course learning (MCL).
  - Three self-supervision courses are designed to alleviate inherent flaws of MLM and balance the label in a multi-perspective way.
  - Two self-correction courses are proposed to bridge the chasm between the two encoders by creating a "correction notebook" for secondary-supervision.
  - A course soups trial is conducted to solve the "tug-of-war" dynamics problem.
- **Results** Our method significantly improves ELECTRA's average performance by 2.8% and 3.2% absolute points respectively on GLUE and SQuAD 2.0 benchmarks, and overshadows recent advanced ELECTRA-style models under the same settings.
- **Resource** The pre-trained MCL model with all evaluation results are available at https://huggingface.co/McmanusChen/MCL-base.

Preliminary & Challenges

- Preliminary

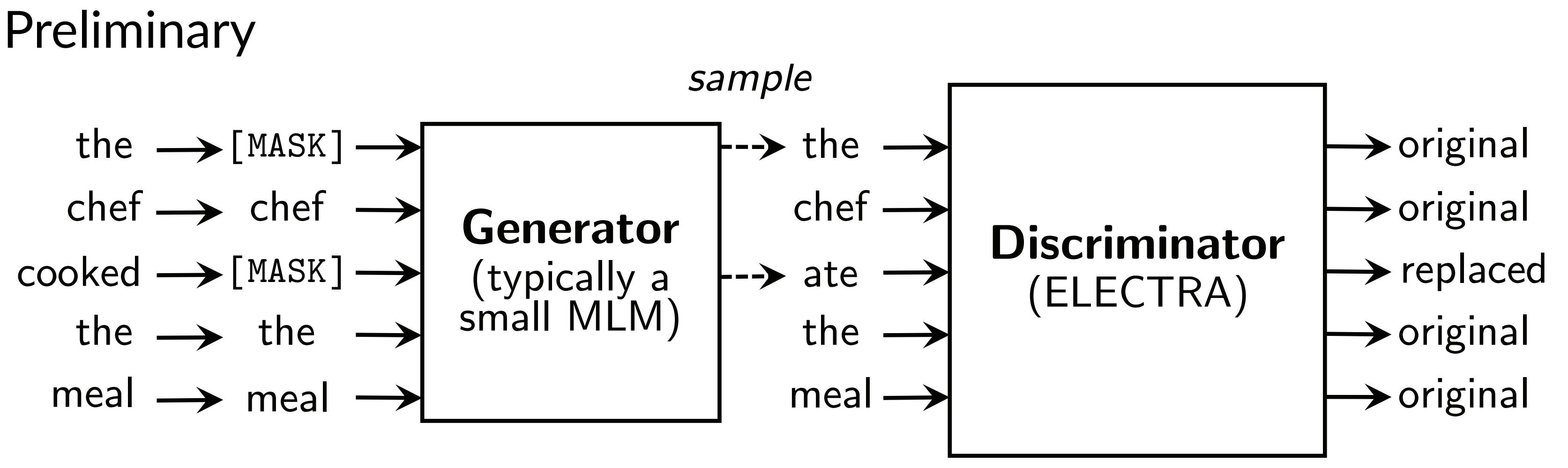


Figure 1. The overall framedwork of ELECTRA (Clark et al., 2020).

- Challenges
  - Biased Learning: G might predict appropriate but not original token on the `[MASK]` position, and such appropriate expression still needs to be judged as substitution by D; the label-imbalance may gradually emerge with the MLM training of G, which could disturb the RTD training of D.
  - Deficient Interaction: there is no explicit feedback loop from D to G, resulting that the pre-training of G is practically dominated by MLM as before.

Multi-perspective Course Learning

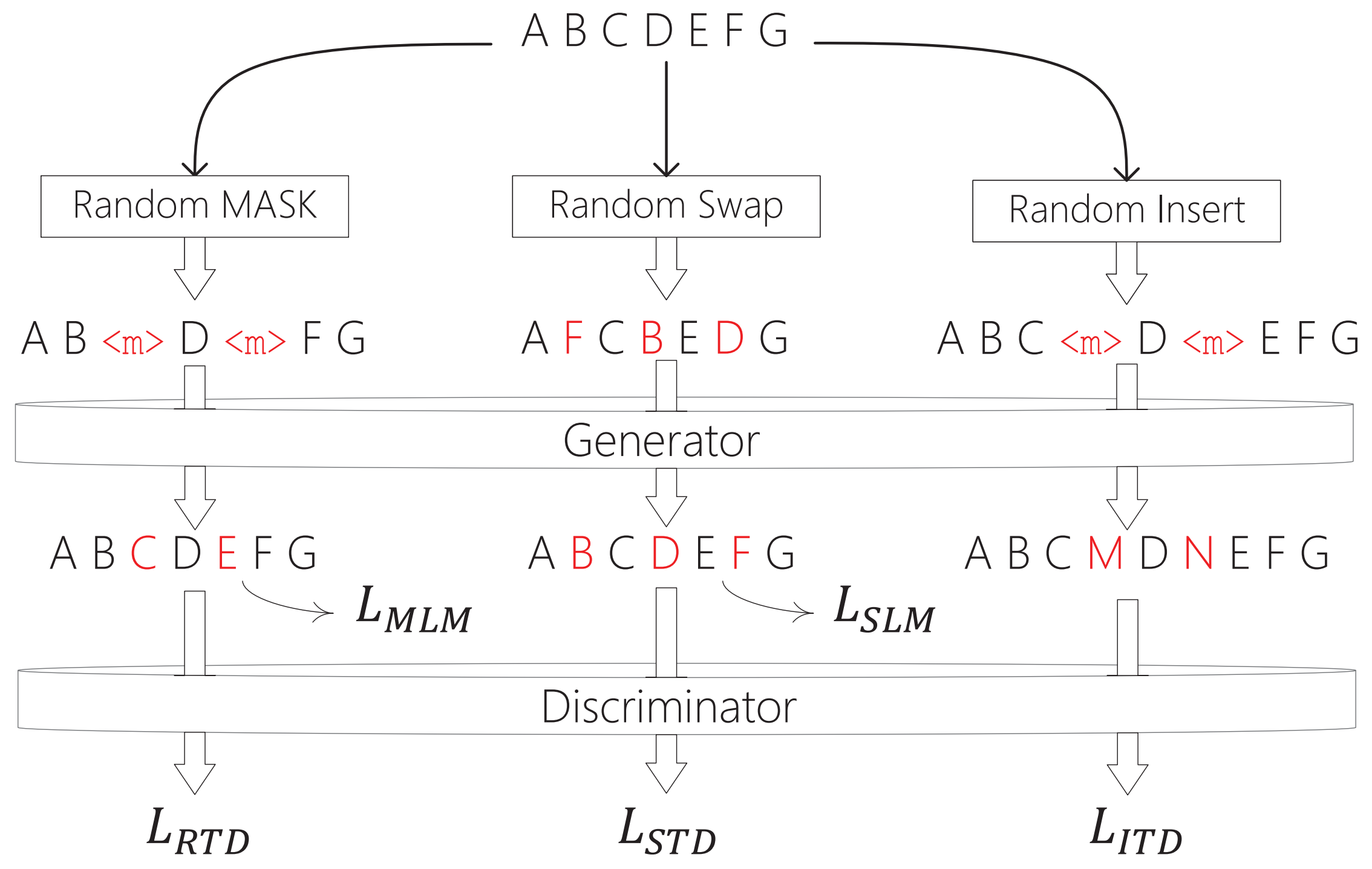
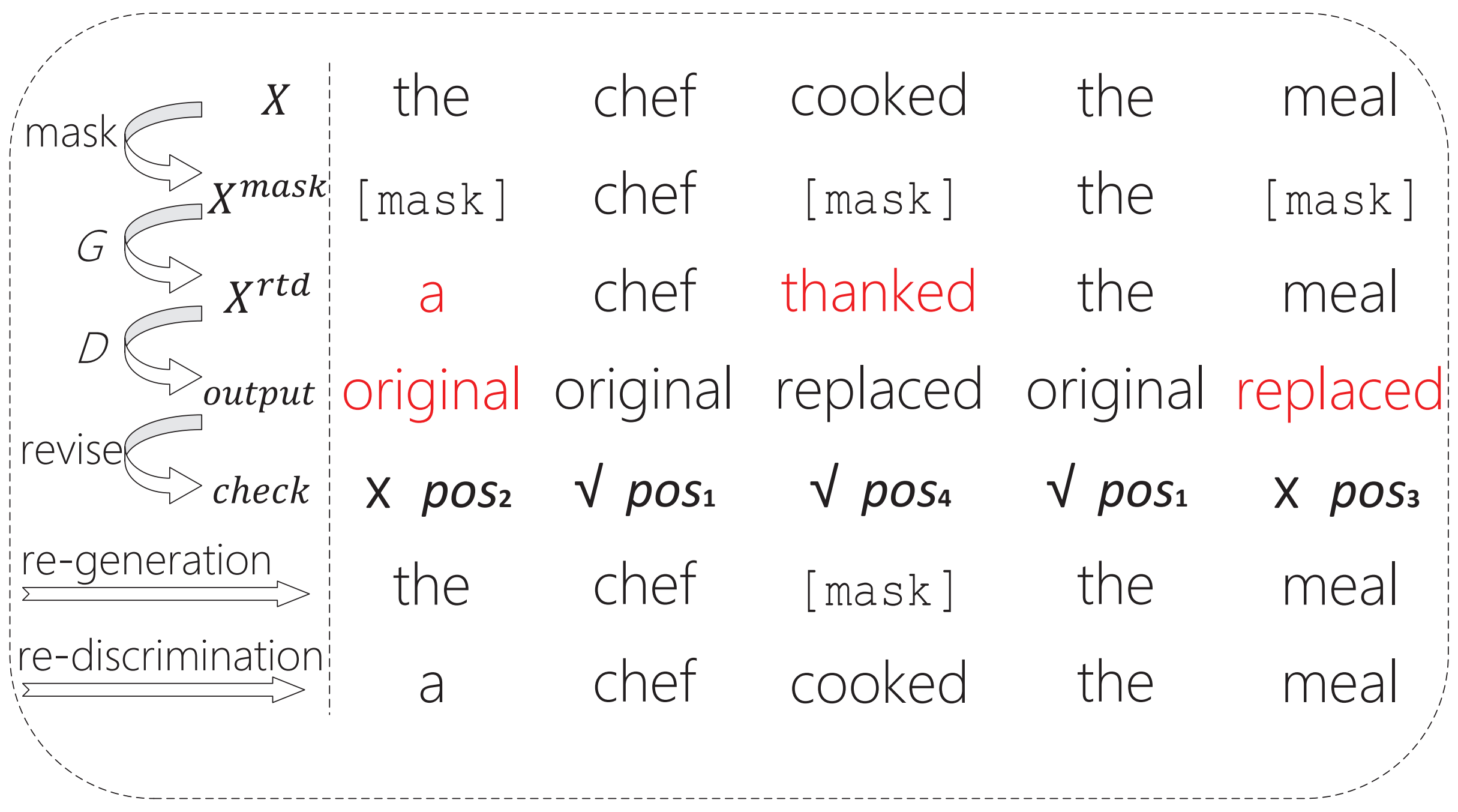


Figure 2. The overall structure of the self-supervision courses. `<m>` denotes `[MASK]`. A capital letter stands for a token and letters in red indicate operated positions.

Predict\Label	original	replaced
original	✓ pos_1	✗ pos_2
replaced	✗ pos_3	✓ pos_4

Table 1. The confusion matrix of output tokens from D. ✓denotes that D makes a correct judgment, conversely ✗ presents the situation of wrong discrimination.



X	the	chef	cooked	the	meal
X^{mask}	`[mask]`	chef	`[mask]`	the	`[mask]`
X^{rtd}	a	chef	thanked	the	meal
output	original	original	replaced	original	replaced
check	x pos_2	√ pos_1	√ pos_4	√ pos_1	x pos_3
re-generation	the	chef	`[mask]`	the	meal
re-discrimination	a	chef	cooked	the	meal

Figure 3. An example for self-correction course of RTD.

- **Self-supervision Course** Replaced token detection (RTD), swapped token detection (STD) and inserted token detection (ITD) are proposed to extend the perspective that models look at sequences.

- **Self-correction Course** An "correction notebook" for G and D (as shown in Table 1) is built by sorting out and analyzing errors, guiding the re-generation and re-discrimination training for secondary-supervision.

Results on GLUE & SQuAD 2.0

Model	GLUE Single Task									SQuAD 2.0	
	MNLI -m/-mm	QQP Acc	QNLI Acc	SST-2 Acc	CoLA MCC	RTE Acc	MRPC Acc	STS-B PCC	AVG	EM	F1
Base Setting: BERT Base Size, Wikipedia + Book Corpus											
BERT	84.5/-	91.3	91.7	93.2	58.9	68.6	87.3	89.5	83.1	73.7	76.3
XLNet	85.8/85.4	-	-	92.7	-	-	-	-	-	78.5	81.3
RoBERTa	85.8/85.5	91.3	92.0	93.7	60.1	68.2	87.3	88.5	83.3	77.7	80.5
DeBERTa	86.3/86.2	-	-	-	-	-	-	-	-	79.3	82.5
TUPE	86.2/86.2	91.3	92.2	93.3	63.6	73.6	89.9	89.2	84.9	-	-
MC-BERT	85.7/85.2	89.7	91.3	92.3	62.1	75.0	86.0	88.0	83.7	-	-
ELECTRA	86.9/86.7	91.9	92.6	93.6	66.2	75.1	88.2	89.7	85.5	79.7	82.6
+HP$_{\text{Loss}}$+Focal	87.0/86.9	91.7	92.7	92.6	66.7	81.3	90.7	91.0	86.7	82.7	85.4
CoCo-LM	**88.5**/88.3	92.0	93.1	93.2	63.9	**84.8**	91.4	90.3	87.2	82.4	85.2
MCL	**88.5/88.5**	**92.2**	**93.4**	**94.1**	**70.8**	84.0	**91.6**	**91.3**	**88.3**	**82.9**	**85.9**
Tiny Setting: A quarter of training flops for ablation study, Wikipedia + Book Corpus											
ELECTRA(*reimplement*)	85.80/85.77	91.63	92.03	92.70	65.49	74.80	87.47	89.02	84.97	79.37	81.31
+STD	86.97/86.97	92.07	92.63	93.30	70.25	82.30	91.27	90.72	87.38	81.73	84.55
+ITD	87.37/87.33	91.87	92.53	93.40	68.45	81.37	90.87	90.52	87.08	81.43	84.20
Self-supervision	87.27/87.33	91.97	92.93	93.03	67.86	82.20	90.27	90.81	87.07	81.87	84.85
+ re-RTD	87.57/87.50	92.07	92.67	92.97	69.80	83.27	91.60	90.71	87.57	81.70	84.48
+ re-STD	87.80/87.77	91.97	92.93	93.33	**71.25**	82.80	91.67	**90.95**	**87.83**	81.81	84.71
MCL	**87.90/87.83**	**92.13**	**93.00**	**93.47**	68.81	**83.03**	**91.67**	90.93	87.64	**82.04**	**84.93**

Table 2. All evaluation results on GLUE and SQuAD 2.0 datasets for comparison. Acc, MCC, PCC, EM, F1 denote accuracy, Matthews correlation, Spearman correlation, Exact-Match and F1 score respectively. Reported results are medians over five random seeds.

Analysis

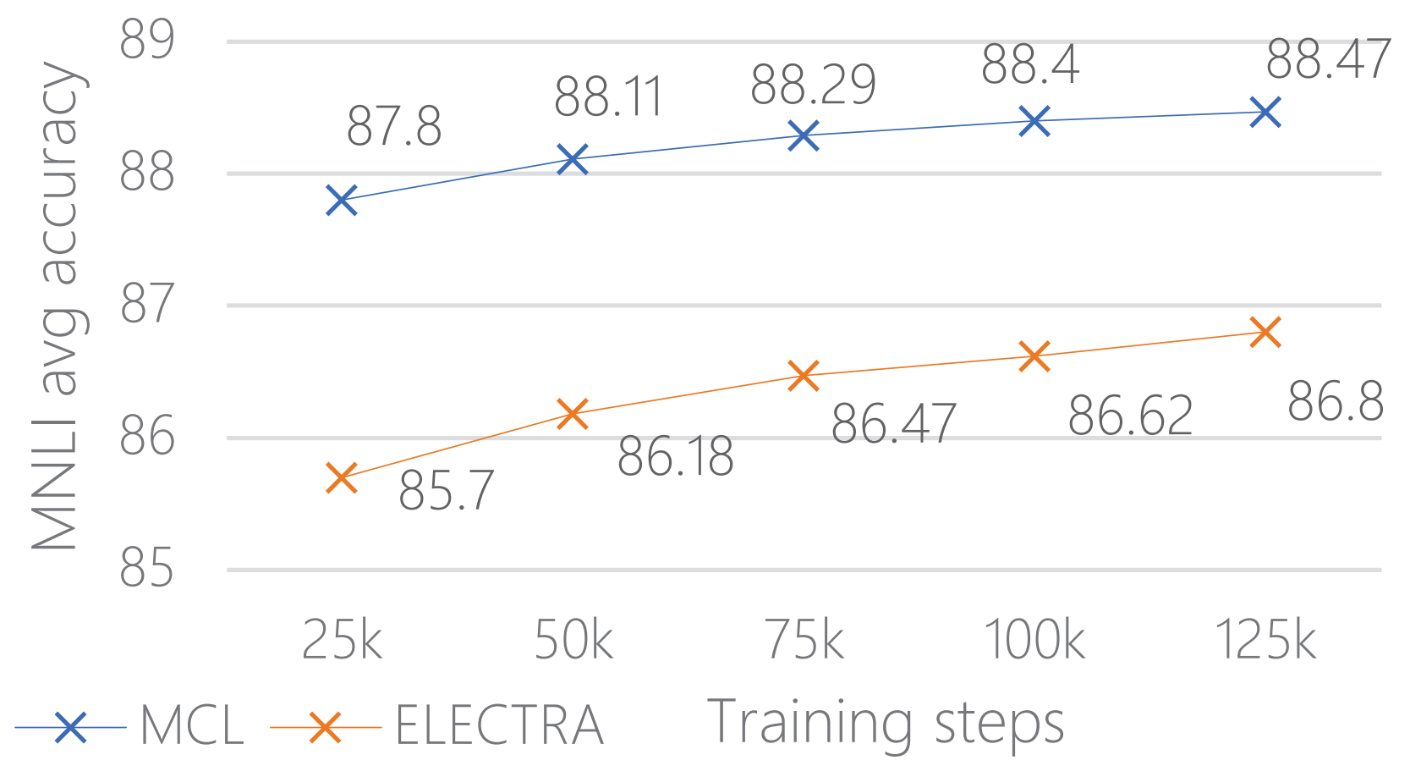


Figure 4. MNLI Comparison of pre-training efficiency. MCL preponderates over ELECTRA baseline on every training node, demonstrating its enormous learning efficiency even on small corpora.

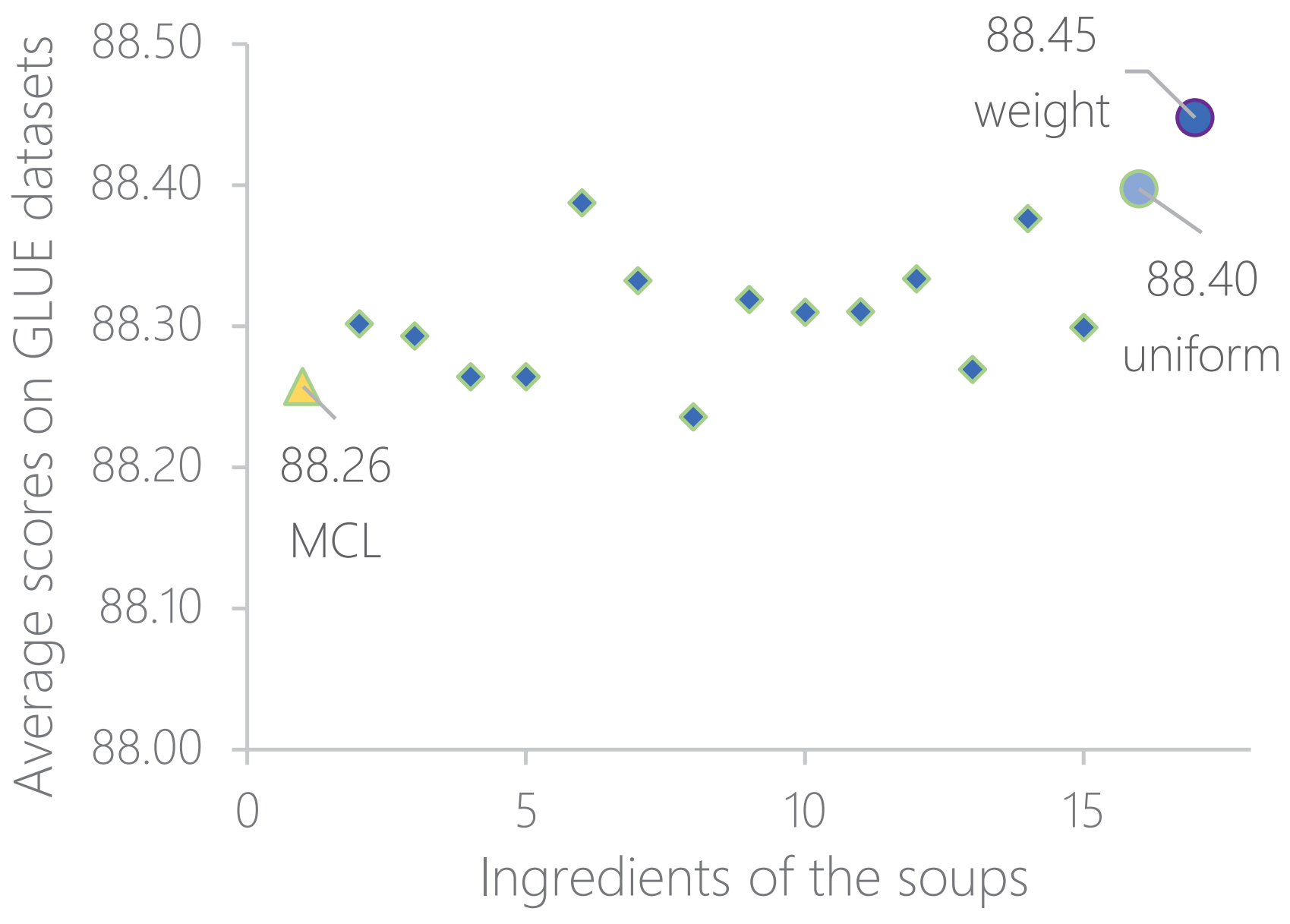


Figure 5. Average GLUE results of the course soups. For ingredients in soups, we arrange all combinations of 4 losses in self-correction courses, training them into 14 single models while retaining the structure of self-supervision courses. Then all ingredients are merged through uniform and weighted integration.

Mail to: beiduo@mail.ustc.edu.cn The 61st Annual Meeting of the Association for Computational Linguistics Paper No.2536