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

Beiduo Chen[1], Shaohan Huang[2], Zihan Zhang[2], Wu Guo[1],
Zhenhua Ling[1], Haizhen Huang[2], Furu Wei[2], Weiwei Deng[2], Qi Zhang[2]

[1]National Engineering Research Center for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China

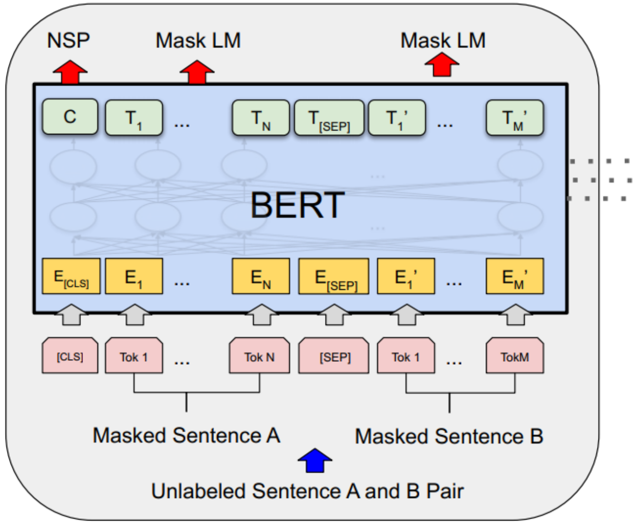
[2]Microsoft Corporation, Beijing, China

Outline

- **Introduction**
- Multi-perspective course learning (MCL)
  - ➢Self-supervision Course
  - ➢Self-correction Course
- Experiments and analyses
- Conclusion

Introduction

- MLM-based Transformer model: BERT, 15% [mask]



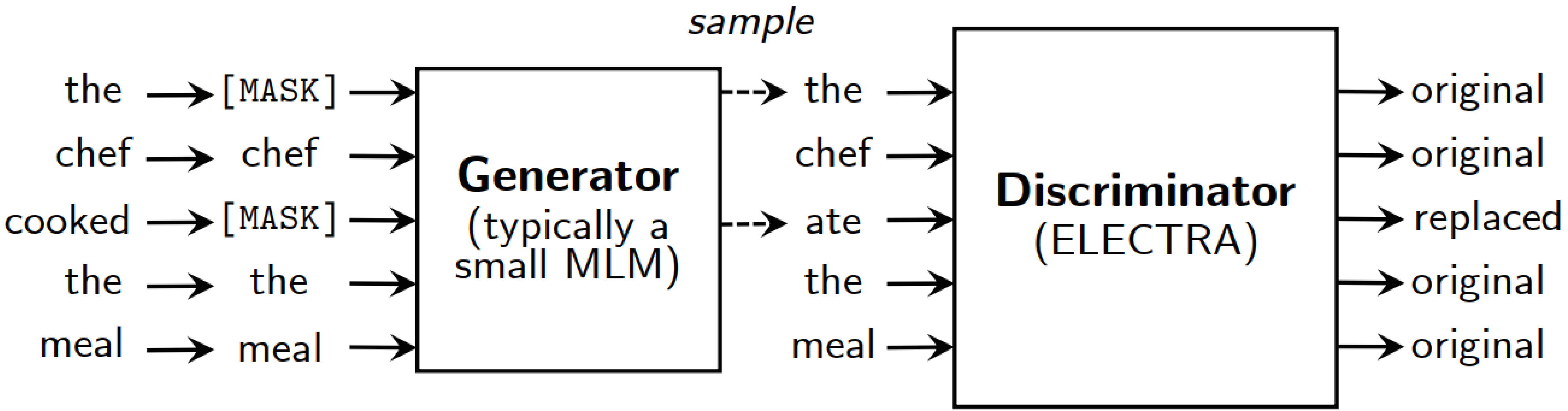
  - ➢ Random corrupt
  - ➢ Unefficient

- → ELECTRA-style framework (Clark et al., 2020)

  - ➢ Challenging ennoising snt
  - ➢ Sample-efficient

ELECTRA(Clark et al., 2020)

- Generator-Discriminator framework: 15%→100% efficiency



- Existing Challenges:
  - Biased Learning: unappropriate questions; label-imbalance
  - Deficient Interaction: no explicit feedback loop from D to G

Outline

- Introduction
- **Multi-perspective course learning (MCL)**
  - ➢Self-supervision Course
  - ➢Self-correction Course
- Experiments and analyses
- Conclusion

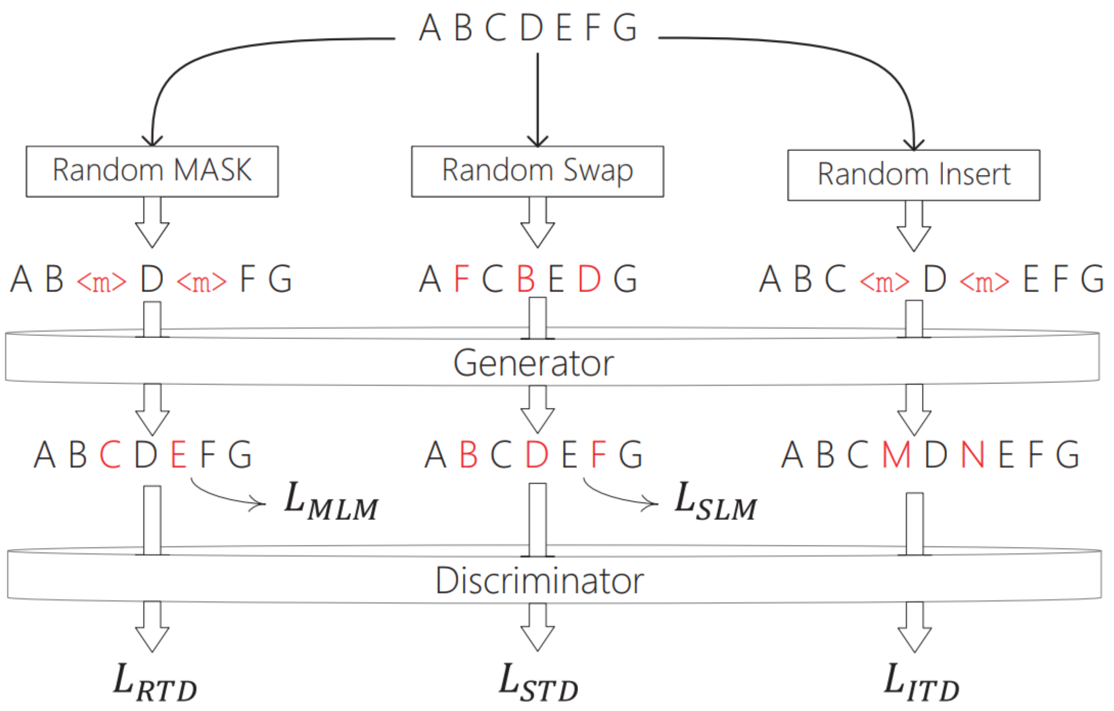
Self-supervision Course

- To extend the perspective that models look at sequences

➢Replaced Token Detection (RTD)

➢Swapped Token Detection (STD)

➢Inserted Token Detection (ITD)



Self-correction Course

- To bridge the chasm between G&D (secondary-supervision)

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

➢Four situations of distinguish results

❑pos1: NaN

❑pos2: re-discriminate

❑pos3: re-discriminate

❑pos4: re-generate

Outline

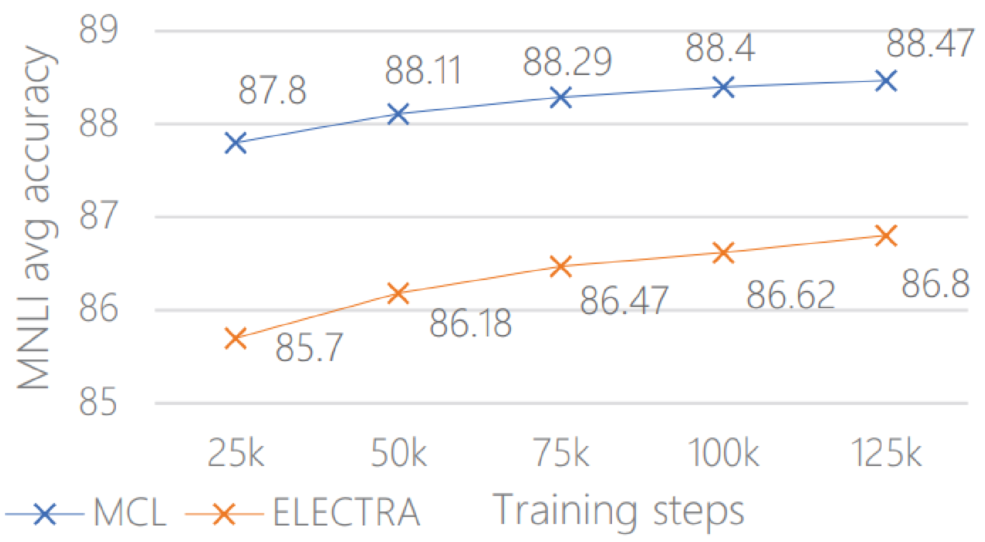
- Introduction
- Multi-perspective course learning (MCL)
  - ➢Self-supervision Course
  - ➢Self-correction Course
- **Experiments and analyses**
- Conclusion

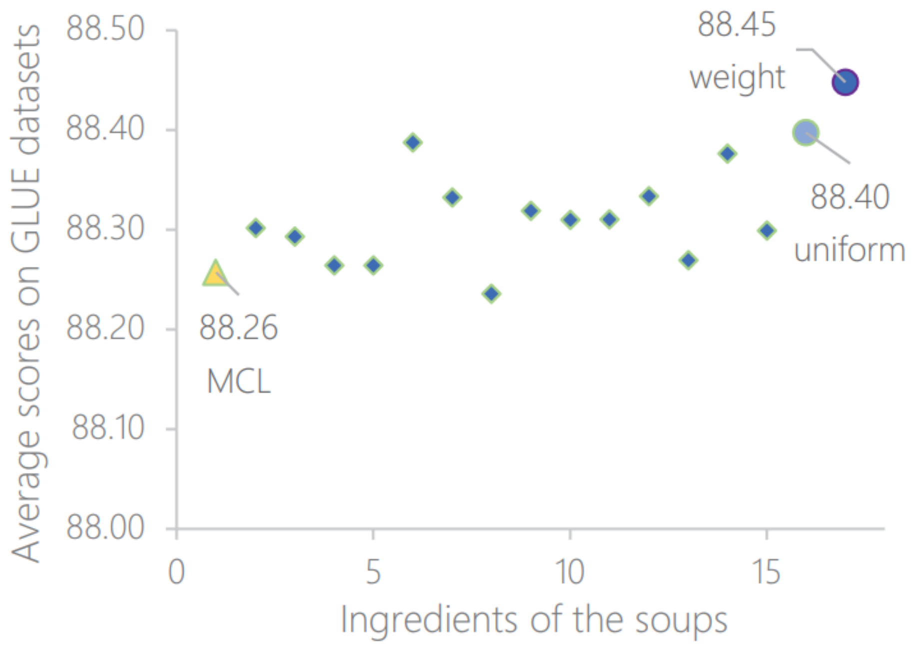
Experiments

Model	GLUE Single Task								
	MNLI -m/-mm	QQP Acc	QNLI Acc	SST-2 Acc	CoLA MCC	RTE Acc	MRPC Acc	STS-B PCC	AVG
Base Setting: BERT Base Size, Wikipedia + Book Corpus									
BERT (Devlin et al., 2019)	84.5/-	91.3	91.7	93.2	58.9	68.6	87.3	89.5	83.1
XLNet (Yang et al., 2019)	85.8/85.4	-	-	92.7	-	-	-	-	-
RoBERTa (Liu et al., 2019)	85.8/85.5	91.3	92.0	93.7	60.1	68.2	87.3	88.5	83.3
DeBERTa (He et al., 2021)	86.3/86.2	-	-	-	-	-	-	-	-
TUPE (Ke et al., 2021)	86.2/86.2	91.3	92.2	93.3	63.6	73.6	89.9	89.2	84.9
MC-BERT (Xu et al., 2020)	85.7/85.2	89.7	91.3	92.3	62.1	75.0	86.0	88.0	83.7
ELECTRA (Clark et al., 2020)	86.9/86.7	91.9	92.6	93.6	66.2	75.1	88.2	89.7	85.5
$+\mathrm{HP}_{\mathrm{Loss}}$+Focal (Hao et al., 2021)	87.0/86.9	91.7	92.7	92.6	66.7	81.3	90.7	91.0	86.7
CoCo-LM (Meng et al., 2021)	**88.5**/88.3	92.0	93.1	93.2	63.9	**84.8**	91.4	90.3	87.2
MCL	**88.5/88.5**	**92.2**	**93.4**	**94.1**	**70.8**	84.0	**91.6**	**91.3**	**88.3**
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
+STD	86.97/86.97	92.07	92.63	93.30	70.25	82.30	91.27	90.72	87.38
+ITD	87.37/87.33	91.87	92.53	93.40	68.45	81.37	90.87	90.52	87.08
Self-supervision	87.27/87.33	91.97	92.93	93.03	67.86	82.20	90.27	90.81	87.07
+ re-RTD	87.57/87.50	92.07	92.67	92.97	69.80	83.27	91.60	90.71	87.57
+ re-STD	87.80/87.77	91.97	92.93	93.33	**71.25**	82.80	91.67	**90.95**	**87.83**
MCL	**87.90/87.83**	**92.13**	**93.00**	**93.47**	68.81	**83.03**	**91.67**	90.93	87.64

Model	SQuAD 2.0	
	EM	F1
Base Setting		
BERT (Devlin et al., 2019)	73.7	76.3
XLNet (Yang et al., 2019)	78.5	81.3
RoBERTa (Liu et al., 2019)	77.7	80.5
DeBERTa (He et al., 2021)	79.3	82.5
ELECTRA (Clark et al., 2020)	79.7	82.6
$+\mathrm{HP}_{\mathrm{Loss}}$+Focal (Hao et al., 2021)	82.7	85.4
CoCo-LM (Meng et al., 2021)	82.4	85.2
MCL	**82.9**	**85.9**
Tiny Setting for ablation study		
ELECTRA(*reimplement*)	79.37	81.31
+STD	81.73	84.55
+ITD	81.43	84.20
Self-supervision	81.87	84.85
+ re-RTD	81.70	84.48
+ re-STD	81.81	84.71
MCL	**82.04**	**84.93**

Analyses





- Sample-efficient Trial
- Course Soups Trial

Outline

- Introduction
- Multi-perspective course learning (MCL)
    - Self-supervision Course
    - Self-correction Course
- Experiments and analyses
- **Conclusion**

Conclusion

- 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.

➢ Paper at

Thanks!

➢ Model at