Segment, Mask, and Predict: Augmenting Chinese Word Segmentation with Self-Supervision

Mieradilijiang Maimaiti[1], Yang Liu[1], Yuanhang Zheng[1], Gang Chen[1], Kaiyu Huang[2], Ji Zhang[3], Huanbo Luan[1], Maosong Sun[1]

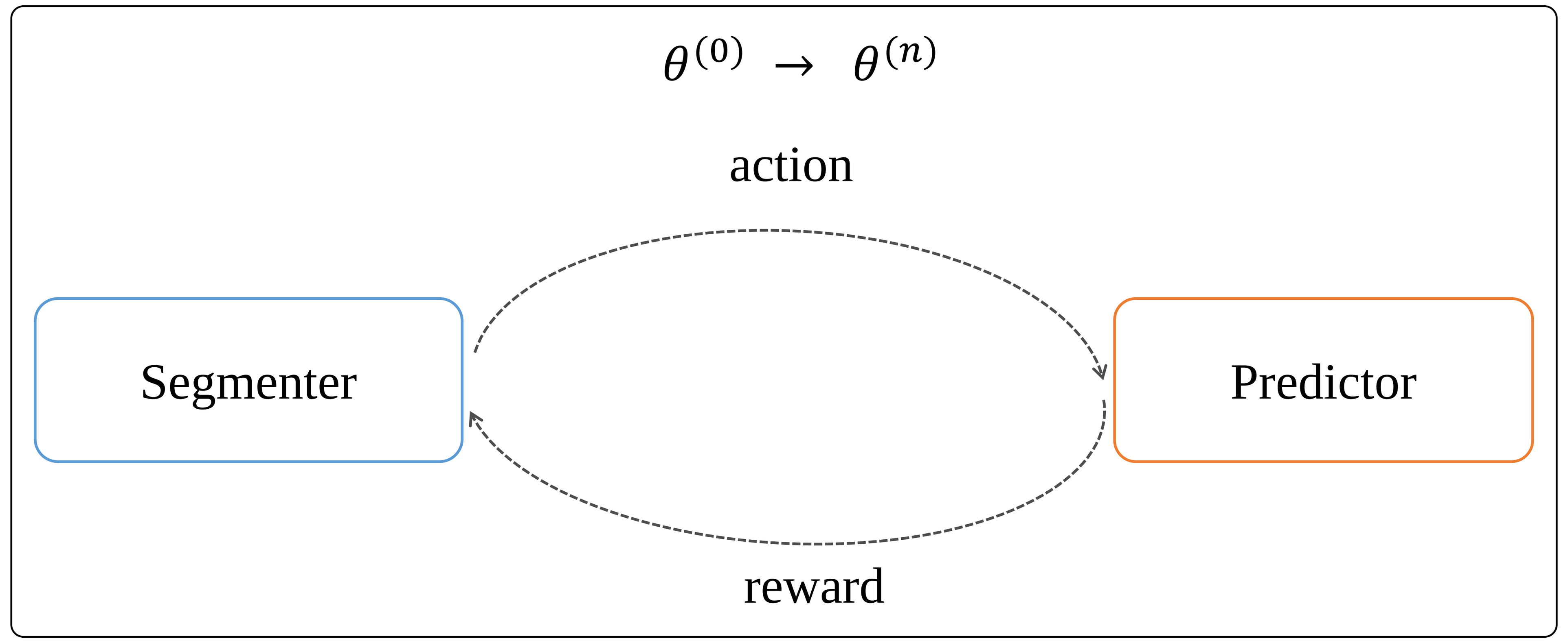
[1]Department of Computer Science and Technology, Tsinghua University [2]School of Computer Science, Dalian University of Technology [3]Alibaba DAMO Academy

E-mail: meadljmm15@mails.tsinghua.edu.cn

Paper Code Blog

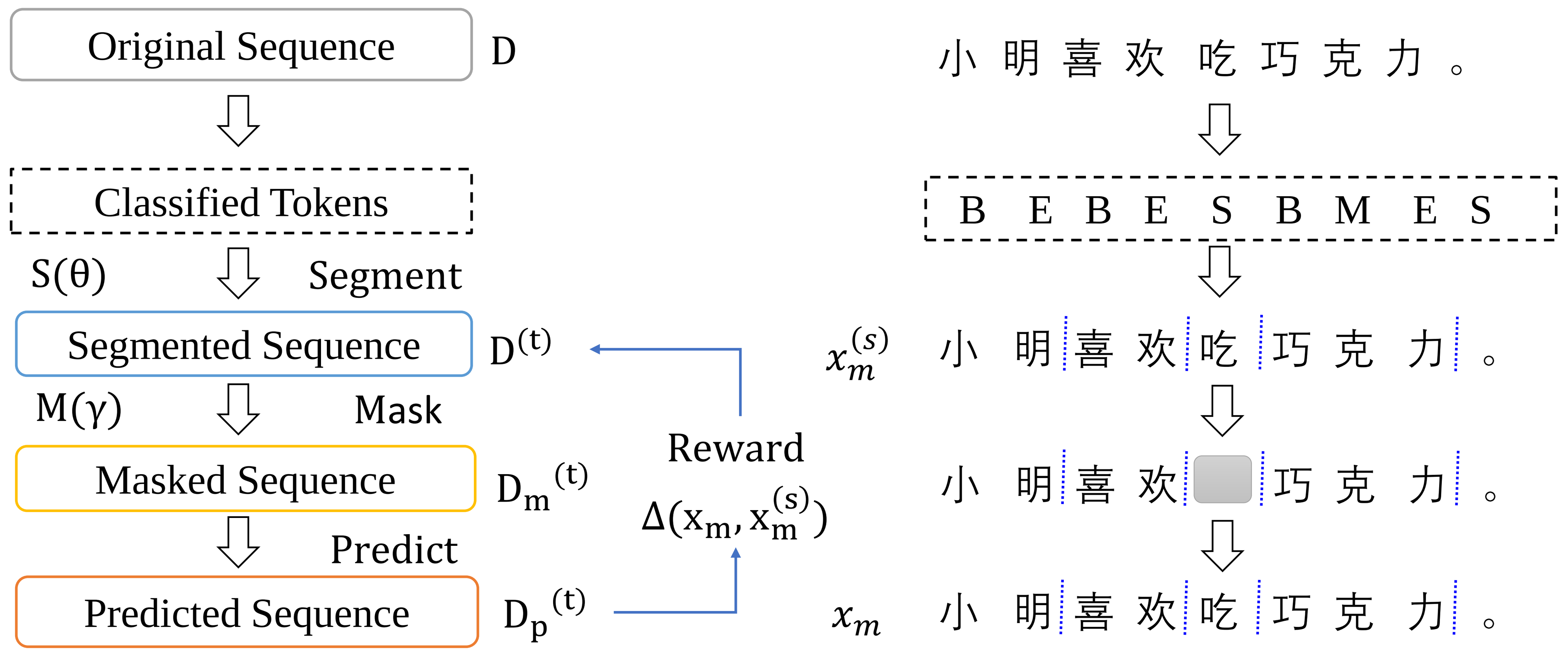
❖ Introduction

- Chinese word segmentation (CWS) is considered an essential task, which will accurately represent semantic information of Chinese NLP tasks.
- Recent SOTA approaches utilize the pre-trained models (PTM) to improve the quality of CWS. However, the CWS methods based on the PTM only utilize the large-scale annotated data to finetune the parameters. It omits much-generated information of the training step.
- Besides, the annotated data has some incorrect labels due to lexical diversity in Chinese, therefore the robustness of methods is quite important for the CWS.
- To address these issues, we propose a self-supervised CWS approach to enhance the performance of CWS model. We exploit the revised masked language model as a predictor to improve the segmentation model, and leverage an improved version of minimum risk training (MRT) to enhance the segmentation.



❖ Methodology

- Model Architecture



- Overall Algorithm

Algorithm 1 Self-supervised Word Segmentation

Input: Original sequence $D = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$.
Output: Original sequence $D_p^{(t)}$.
1: Train Mask-Predictor $M(\gamma)$ based on D.
2: Train Segmenter $S(\theta^{(o)})$ based on D.
3: Employ $S(\theta^{(o)})$ to segment D and achieve segmented sequence $D^{(t)}$.
4: Mask $D^{(t)}$ to obtain the masked sequence $D_m^{(t)}$ with the strategy.
5: Exploit $M(\gamma)$ to achieve predicted sequence $D_p^{(t)}$ based on $D^{(t)}$.
6: Calculate the accuracy by comparing $D_p^{(t)}$ and $D^{(t)}$ as a reward.
7: Update the $S(\theta^{(o)})$ to $S(\theta^{(n)})$.

- Revised MLM as Predictor

Segged Seq.	小明 喜欢 吃 巧克力 。
Masked Input	[M] [M] 喜 欢 吃 巧 克 力 。 小 明 [M] [M] 吃 巧 克 力 。 小 明 喜 欢 [M] 巧 克 力 。 小 明 喜 欢 吃 [M] [M] 力 。 小 明 喜 欢 吃 巧 [M] [M] 。 小 明 喜 欢 吃 巧 克 力 [M]

- Training Procedure with Improved MRT

$$J(\theta) = \sum_{\mathbf{x} \in \mathbf{X}} \left(\sum_{\mathbf{y} \in S(\mathbf{x})} Q(\mathbf{y}|\mathbf{x}; \theta, \alpha) q(\mathbf{y}, \mathbf{x}) - \lambda \sum_{\mathbf{y}' \in S(\mathbf{x})} P(\mathbf{y}'|\mathbf{x}; \theta)^{\alpha} \right)$$

❖ Experiments

- Results of Single Criterion Learning

Methods	SIGHAN05				SIGHAN08		OTHER		
	MSRA	PKU	AS	CITYU	CTB	SXU	CNC	UDC	ZX
Chen et al. (2017)	95.84	93.30	94.20	94.07	95.30	95.17	—	—	—
Zhou et al. (2017)	97.80	96.00	—	—	96.20	—	—	—	—
Yang et al. (2017)	97.50	96.30	95.70	96.90	96.20	—	—	—	—
He et al. (2018)	97.29	95.22	94.90	94.51	95.21	95.78	97.11	93.98	95.57
Gong et al. (2019)	96.46	95.74	94.51	93.71	97.09	95.57	—	—	—
LSTM+BEAM	97.10	95.80	95.30	95.60	96.10	95.95	96.10	96.20	96.30
LSTM+CRF	98.10	96.10	96.00	96.80	96.30	96.55	96.61	96.00	96.40
BERT	96.91	95.34	96.47	97.10	97.27	96.40	96.66	97.23	96.49
SELFATT+SOFT	97.60	95.50	95.70	96.40	97.28	96.60	96.88	97.12	96.50
BERT+LTL	97.53	96.23	97.03	97.63	97.34	96.65	96.89	97.51	96.72
Ours	**98.12**	**96.24**	**97.30**	**97.83**	**97.45**	**96.97**	**97.25**	**97.74**	**96.82**

- Results of Multiple Criteria Learning

Methods	SIGHAN05				SIGHAN08		OTHER		
	MSRA	PKU	AS	CITYU	CTB	SXU	CNC	UDC	ZX
Chen et al. (2017)	96.04	94.32	94.64	95.55	96.18	96.04	—	—	—
He et al. (2018)	97.35	95.78	95.47	95.60	95.84	96.49	97.00	94.44	95.72
Gong et al. (2019)	97.78	96.15	95.22	96.22	97.26	97.25	—	—	—
BERT	97.22	96.06	97.07	97.39	97.36	96.81	96.71	97.48	96.60
BERT+LTL	96.67	96.30	97.16	97.72	97.38	96.90	97.10	97.61	96.81
Ours	**98.19**	**96.32**	**97.43**	**97.80**	**97.66**	**97.03**	**97.34**	**98.25**	**97.08**

- Results on Noisy Datasets

Methods	SIGHAN05				SIGHAN08		OTHER		
	MSRA	PKU	AS	CITYU	CTB	SXU	CNC	UDC	ZX
LSTM+BEAM	96.86	95.70	95.17	95.35	95.89	95.83	95.89	96.07	96.18
LSTM+CRF	97.89	95.89	95.88	96.67	96.19	96.47	96.49	95.85	96.25
BERT	96.78	95.20	96.28	97.01	97.14	96.24	96.51	97.11	96.30
SELFATT+SOFT	97.47	95.40	95.57	96.29	97.16	96.49	96.61	97.08	96.33
BERT+LTL	97.42	96.15	96.76	97.52	97.27	96.55	96.69	97.40	96.53
Ours	**97.93**	**96.18**	**97.12**	**97.68**	**97.32**	**96.83**	**97.12**	**97.63**	**96.67**