



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

**Mieradilijiang Maimaiti**<sup>1</sup>, Yang Liu<sup>1</sup>, Yuanhang Zheng<sup>1</sup>, Gang Chen<sup>1</sup>  
Kaiyu Huang<sup>2</sup>, Ji Zhang<sup>3</sup>, Huanbo Luan<sup>1</sup>, and Maosong Sun<sup>1</sup>

(Hangzhou, 2011.11.18)



# Outline

- Chinese Word Segmentation
- Background & Significance
- Challenges & Motivation
- Methodology
- Experiment & Results
- Conclusion & Future Work

# Chinese Word Segmentation



# Conception

- Much like **sentences** are composed of **words**, words themselves are composed of **smaller units**.
- Chinese sentences consist of chars which is the smallest unit.

The diagram illustrates the segmentation of the English word "Unquestionably" and its Chinese translation "毫无疑问的".

**English Word Segmentation:**

- surface form:** Unquestionably
- underlying form:** Unquestionably
- segmentation:** The word is segmented into four parts: "un" (prefix), "question" (stem), "able" (suffix), and "ly" (suffix). Dashed lines connect the labels to the corresponding segments.

**Chinese Translation:**

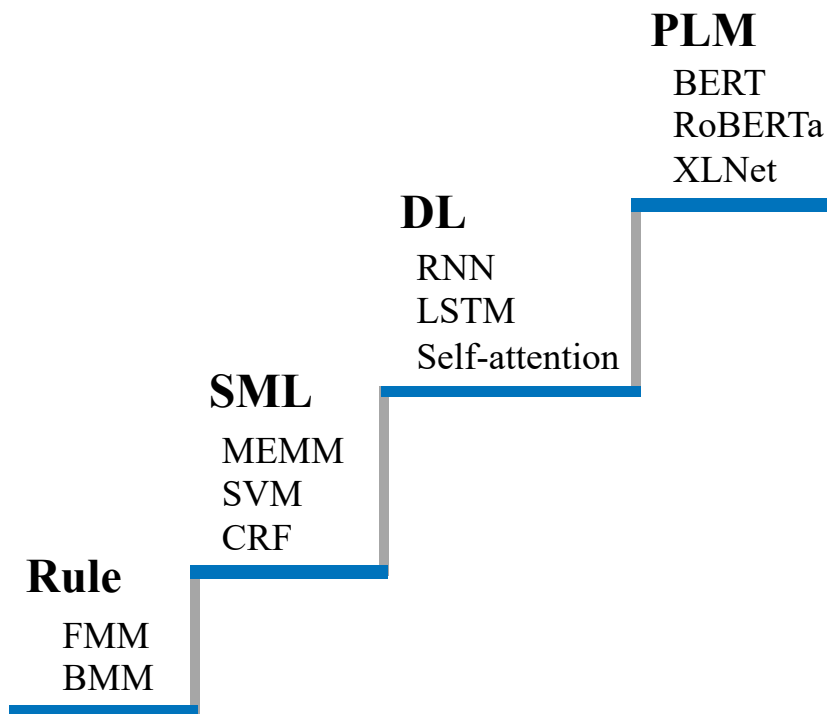
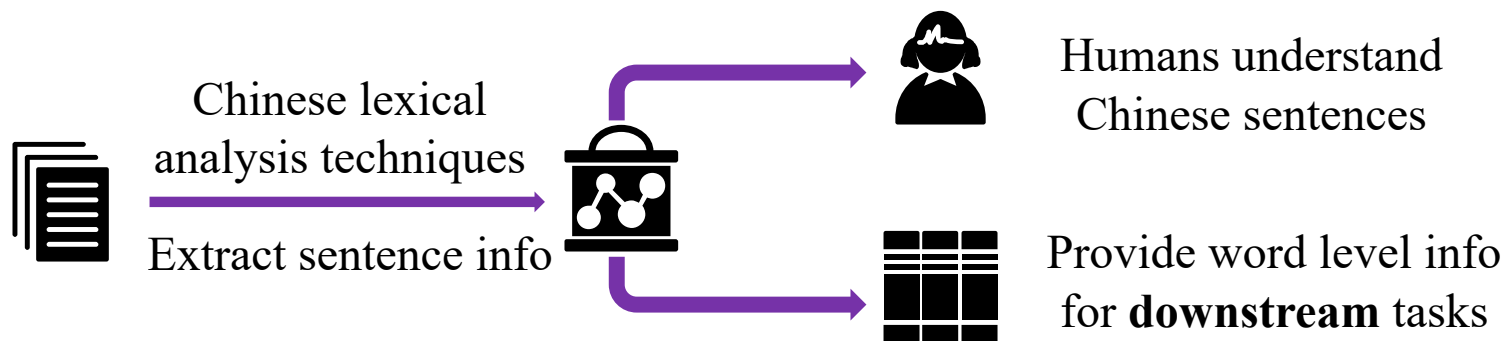
- Original:** 毫无疑问的
- segmentation:** The translation is segmented into three parts: "毫无" (prefix), "疑问" (stem), and "的" (suffix). Dashed lines connect the labels to the corresponding segments.



# Outline

- Chinese Word Segmentation
- Background & Significance
- Challenges & Motivation
- Methodology
- Experiment & Results
- Conclusion & Future Work

# Background



# Significance



## Does it make sense?

- Application value --- MT, IR, NER, NLU, QA...

## Low-Resource Languages NMT

## Cross-Lingual Information Retrieval



# 清华大学跨语言信息检索系统

搜索

实现中华民族伟大复兴

جميىتى ئىزدەپ تاپقان ئۇچۇر: 7872. ئشلەنگەن ۋاقىت: 267 مىكرو سىكۇنت

باش شۇجى شى جىنپىڭ تەيۋەندىكى ھەر ساھە تەربىيىلىرى بىلەن كۆرۈشتى. تەڭرىتاغ تورى  
 ئالاقىدار، تىنچلىقنى قەدەرلەپ، تەرەققىياتنى بىرلىكتە پىلانلاپ، بىر نىيەتتە ئىتتىپاقلىشىپ **جۇڭخۇا مىللەتلىرىنىڭ تۇلۇغ كۆڭلىنىنىڭ**  
**ئىمەلگە ئاشۇش ئۇنۋانى** ...  
[http://ey.ts.cn/topic/kangri/2015\\_09/01/content\\_449260.htm](http://ey.ts.cn/topic/kangri/2015_09/01/content_449260.htm) 2017\_02\_28 22:52:34

خاتالىق مۇساپە، يولاتەك پاكىت- تەڭرىتاغ تورى  
 ئىش قانچىنى، ھازىر شىنجاڭنىڭ تەرەققىياتى يېڭى تارىخىي باشلاندى ئۇنىڭدا تۇرۇۋاتىدۇ. مەملىكەت خەلقى بىلەن بىللە، **جۇڭخۇا**  
**مىللەتلىرىنىڭ تۇلۇغ كۆڭلىنىنىڭ ئىمەلگە**  
 شاتلىق يىپەك يولىنىڭ يېڭى سەھىپىسىنى بىرلىكتە ئاچايلى- تەڭرىتاغ تورى  
 قاتارلىق ئەۋزەللىكلەرنى ئىمەلگە ھەمكارلاشتى، ئىقتىسادنى سىجىل **ئاشۇش**، خىزمەتچىلەرنى ئورتاق ئاقايلى تۇرۇش ئەۋزەللىكىگە  
 ئايلاندۇردى. يىپەك يولى ئىقتىساد بەلبېغىنى ئورتاق قۇرۇش **جۇڭخۇا مىللەتلىرىنىڭ تۇلۇغ** ...  
[http://ey.ts.cn/shuanti/2014\\_09/04/content\\_372758.htm](http://ey.ts.cn/shuanti/2014_09/04/content_372758.htm) 2018\_07\_05 22:29:24

ئالدىنقىلارغا ۋارىسلىق قىلىپ، كېيىنكىلەرگە يول ئېچىش- تەڭرىتاغ تورى  
**جۇڭخۇا مىللەتلىرىنىڭ تۇلۇغ كۆڭلىنىنىڭ ئىمەلگە ئاشۇرۇشا** مۇناسىۋەتلىك، دېمەك كۆرسەتتى. مەركەز ھەر **مىللەت** نەقلىگە  
**جۇڭخۇا مىللەتلىرىنىڭ تۇلۇغ كۆڭلىنىنىڭ ئىمەلگە ئاشۇرۇش ئۇنۋانى** ...  
[http://ey.ts.cn/topic/6dzhounian/2015\\_09/10/content\\_451221.htm](http://ey.ts.cn/topic/6dzhounian/2015_09/10/content_451221.htm) 2018\_07\_05 22:33:43

خاتالىق مۇساپە، يولاتەك پاكىت- تەڭرىتاغ تورى  
 ئىش قانچىنىڭ تەرەققىياتى يېڭى تارىخىي باشلاندى ئۇنىڭدا تۇرۇۋاتىدۇ. مەملىكەت خەلقى بىلەن بىللە، **جۇڭخۇا مىللەتلىرىنىڭ**  
**تۇلۇغ كۆڭلىنىنىڭ ئىمەلگە ئاشۇرۇش ئۇنۋانى** ...  
[http://ey.ts.cn/topic/6dzhounian/2015\\_09/26/content\\_450155.htm](http://ey.ts.cn/topic/6dzhounian/2015_09/26/content_450155.htm) 2018\_07\_06 09:07:16

جۇڭخۇا خەلق سىياسىي مەسلىھەت كېڭەش مەملىكەتلىك 12- نۆۋەتلىك كومىتېتى 4- يىغىنىنىڭ ...  
 مەملىكەتلىرىنىڭ **تۇلۇغ كۆڭلىنىنىڭ ئىمەلگە ئاشۇرۇش كېرەك**. دۆلىتىمىزنىڭ سىرتقا قارىتىلغان خىزمىتىدىكى ئومۇمىي ئورۇنلاشتۇرۇش سىياسىسىدا  
 ئاساسەن سىرت بىلەن بولغان دوستلار ئالاقىدا تۇرۇش ...  
[http://ey.ts.cn/2016lianghui/2016\\_03/15/content\\_509966.htm](http://ey.ts.cn/2016lianghui/2016_03/15/content_509966.htm) 2017\_05\_22 22:23:23

جۇڭخۇا خەلق سىياسىي مەسلىھەت كېڭەش مەملىكەتلىك 12- نۆۋەتلىك كومىتېتى 4- يىغىنىنىڭ ...



Does it make sense?

- Academic value

CWS for NMT

Segmentation Method	BLEU (Zh – En)
CHAR	21.16
TEACHER	23.51
CRF	23.37
CONPRUNE	<b>23.73</b>

(Huang et al., 2021)

CWS for Name Entity Recognition

Segmentation Method	NR	NP	NT
CHAR	89.50	88.00	86.40
TEACHER	89.70	87.50	86.20
CRF	90.70	88.00	87.70
CONPRUNE	<b>91.50</b>	<b>88.40</b>	<b>87.70</b>

(Huang et al., 2021)



# Outline

- Chinese Word Segmentation
- Background & Significance
- Challenges & Motivation
- Methodology
- Experiment & Results
- Conclusion & Future Work

# Challenges



## Main challenges

- Annotation inconsistency
  - 操作系统 (operating system) VS. 操作 (operating) / 系统 (system)
  - eight times six times
- Word boundary detection
  - 犯罪(crime) / 案(case) 走私案 (smuggling case)

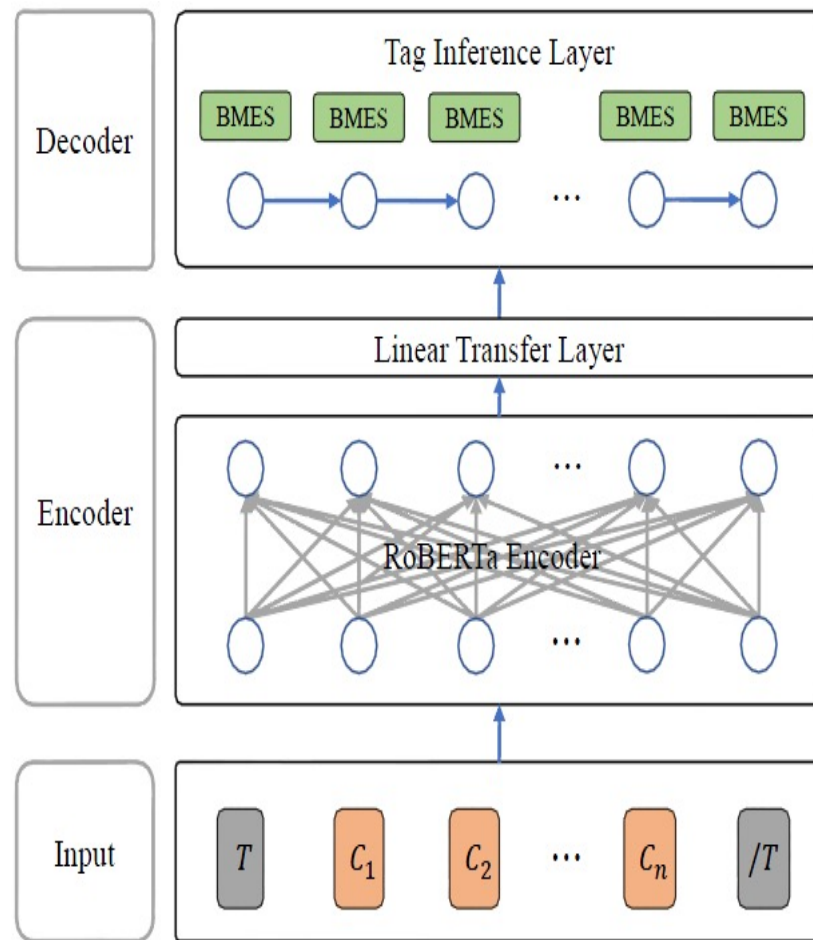
### Same sentences in different corpus

Corpus	Zhang	Xiao	Fan	attend	a tournament	
PKU	张	小凡		参加	比武	大会
MSRA	张小凡			参加	比武大会	
Zhuxian	张小凡			参加	比武	大会



## Main challenges

- Complex architecture
  - Computational cost
  - Memory consumption
  - RoBERTa
  - GPU
    - 1080 or TITAN
    - 12G memory ❌
    - 3090
    - 24G memory ✅
- Poor robustness



(Huang et al., 2020)

# Motivation



## What motivates us?

- Steady model
  - Word, phrase and sentence level inconsistency
- Cheaper computational resource
  - Lower GPU memory
- Better robustness
  - Different corpora
  - Different domain

(Huang et al., 2020)

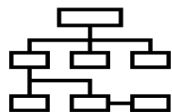




# Outline

- Chinese Word Segmentation
- Background & Significance
- Challenges & Motivation
- Methodology
- Experiment & Results
- Conclusion & Future Work

# Methodology



## General architecture of CWS

- Input sequence (Char level)

$$X = \{x_1, \dots, x_n\}; Y^* = \{y_1^*, \dots, y_n^*\}$$

$$Y^* = \arg \max_{Y \in \mathcal{L}^n} p(Y|X)$$

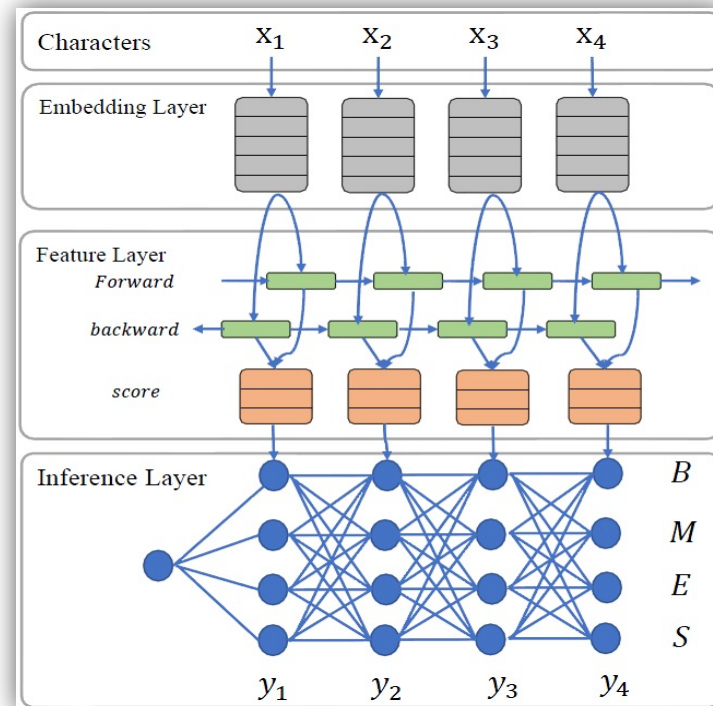
$$\mathcal{L} = \{B, M, E, S\}$$

- Vector representation

- Mapping  $x_i$  into  $\mathbf{e}_{x_i} \in \mathbb{R}^{d_e}$

- Feature extraction

$$\begin{aligned} \mathbf{h}_i &= \vec{\mathbf{h}}_i \oplus \overleftarrow{\mathbf{h}}_i \\ &= \text{Bi-LSTM}(\mathbf{e}_{x_i}, \vec{\mathbf{h}}_{i-1}, \overleftarrow{\mathbf{h}}_{i+1}, \theta) \end{aligned}$$



(Chen et al., 2017)

- Output (CRF 4 labels)

$$p(Y|X) = \frac{\Psi(Y|X)}{\sum_{Y' \in \mathcal{L}^n} \Psi(Y'|X)}$$

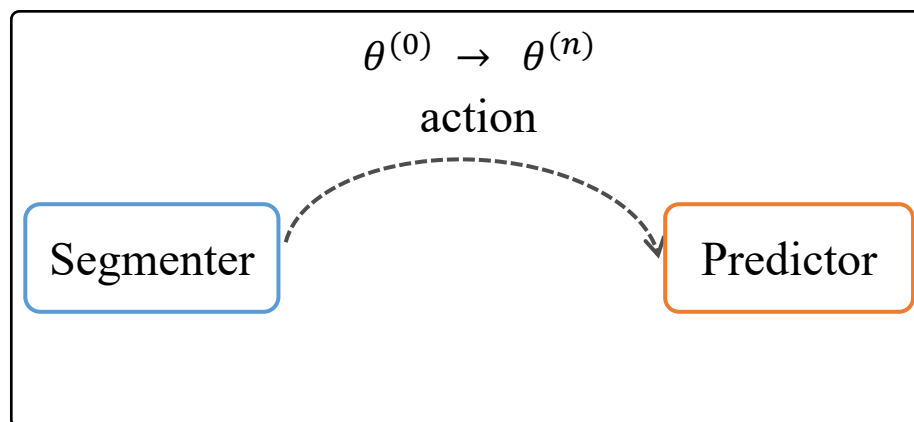


# Self-supervised word segmentation model

Segmenter

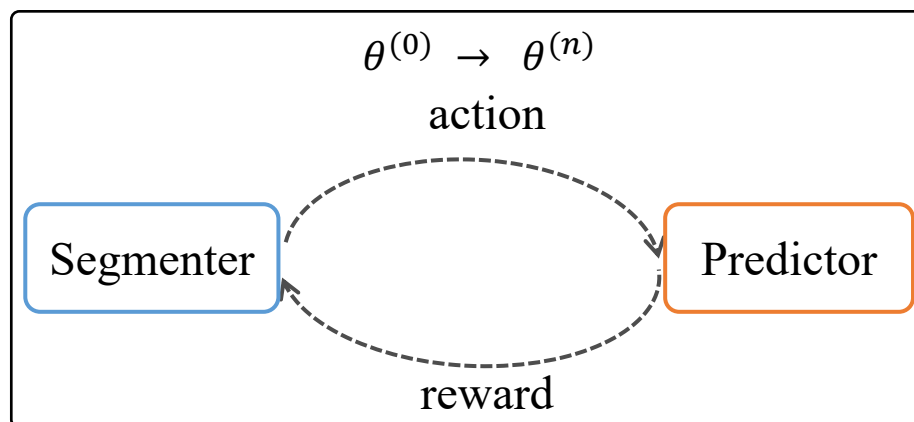



# Self-supervised word segmentation model





# Self-supervised word segmentation model





Segementer

The diagram shows a large rectangular box with a light yellow dotted background. Inside this box, on the left side, is a smaller rounded rectangular box with a blue border. The word "Segementer" is written inside this smaller box.



## How does it work?

- Input sequence

$$\begin{aligned} q(\mathbf{y}|\mathbf{x}) &= \mathbb{E}_{\mathbf{x}_m|\mathbf{x}_o^{(s)}, \mathbf{y}; \gamma} \left[ \Delta \left( \mathbf{x}_m, \mathbf{x}_m^{(s)} \right) \right] \\ &= \sum_{\mathbf{x}_m \in M(\mathbf{x}, \mathbf{y})} P \left( \mathbf{x}_m | \mathbf{x}_o^{(s)}; \gamma \right) \Delta \left( \mathbf{x}_m, \mathbf{x}_m^{(s)} \right) \end{aligned}$$

- $\mathbf{x}$  input seq,  $\mathbf{y}$  label seq;
- $M(\mathbf{x}, \mathbf{y})$  all the legal masking of  $\mathbf{x}$  when seg result is  $\mathbf{y}$ .
- $\mathbf{x}_m$  predicted result,  $\mathbf{x}_m^{(s)}$  ground truth of masked part,  $\mathbf{x}_o^{(s)}$  non-masked part of MLM.

$$\Delta \left( \mathbf{x}_m, \mathbf{x}_m^{(s)} \right) = 1 - \text{sim} \left( \mathbf{x}_m, \mathbf{x}_m^{(s)} \right)$$





## Revised masking strategy

**All the legal masked sequence when Mask count = 2**

Segmented sequence	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]



## How to optimize the model?

- Training step is similar to MRT (Shen et al., 2016)

$$J(\theta) = \sum_{\mathbf{x} \in \mathbf{X}} \mathbb{E}_{\mathbf{y}|\mathbf{x}; \theta} [q(\mathbf{y}|\mathbf{x})] = \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in Y(\mathbf{x})} P(\mathbf{y}|\mathbf{x}; \theta) q(\mathbf{y}|\mathbf{x})$$

- $Y(\mathbf{x})$  is the set of all the possible segmentation results.
- Hard to calculate the cost, need to sample a sub-set  $S(\mathbf{x})$ .

$$Q(\mathbf{y}|\mathbf{x}; \theta, \alpha) = \frac{P(\mathbf{y}|\mathbf{x}; \theta)^\alpha}{\sum_{\mathbf{y}' \in S(\mathbf{x})} P(\mathbf{y}'|\mathbf{x}; \theta)^\alpha}$$

- Final training procedure with improved MRT.

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



## Model Architecture

Original Sequence <sup>D</sup>



Classified Tokens

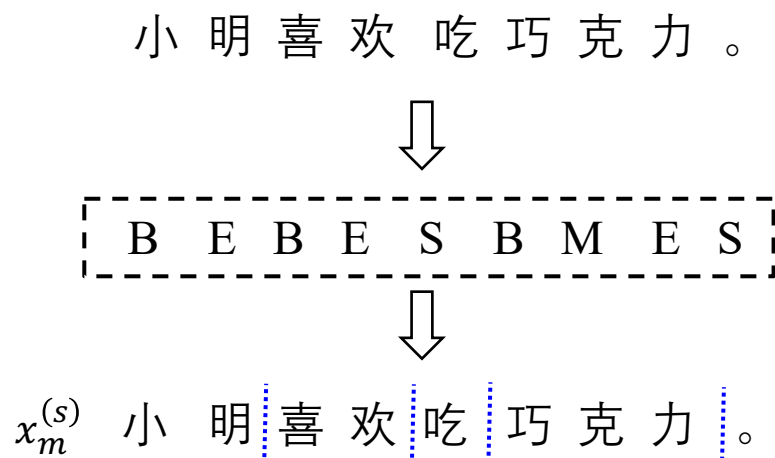
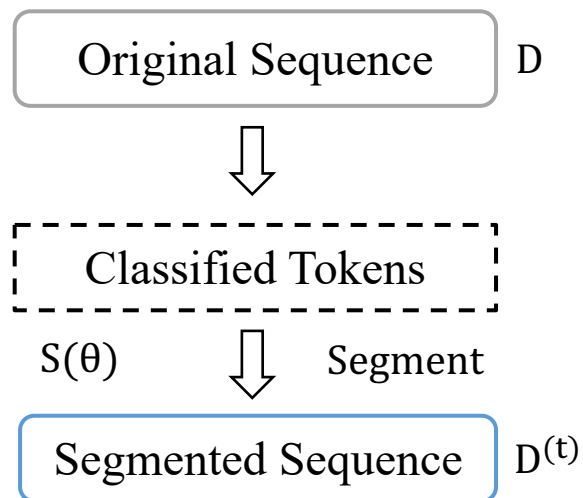
小 明 喜 欢 吃 巧 克 力 。



B E B E S B M E S

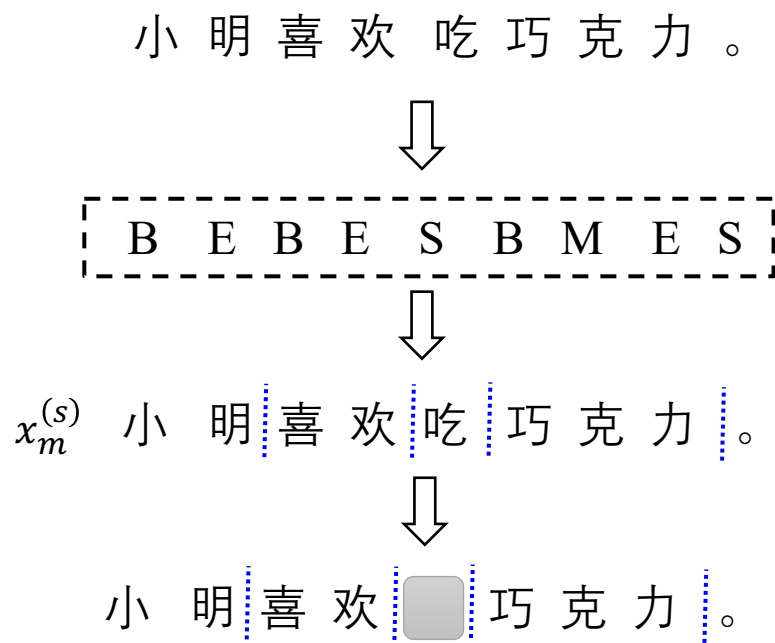
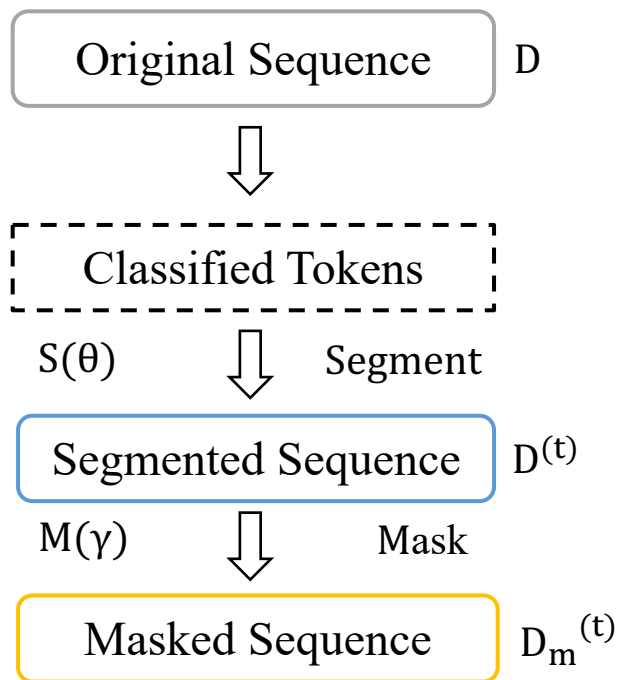


## Model Architecture



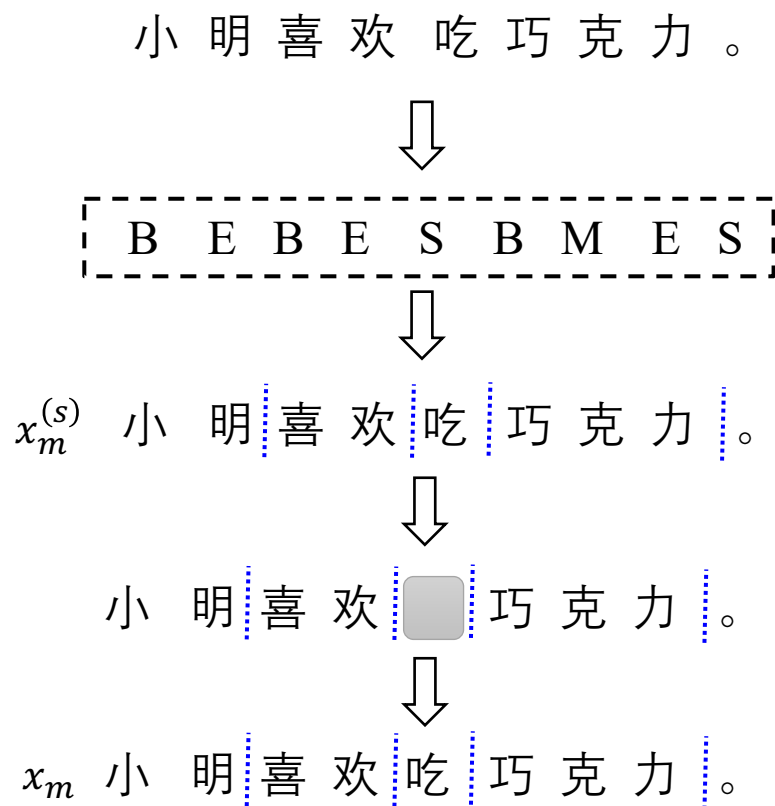
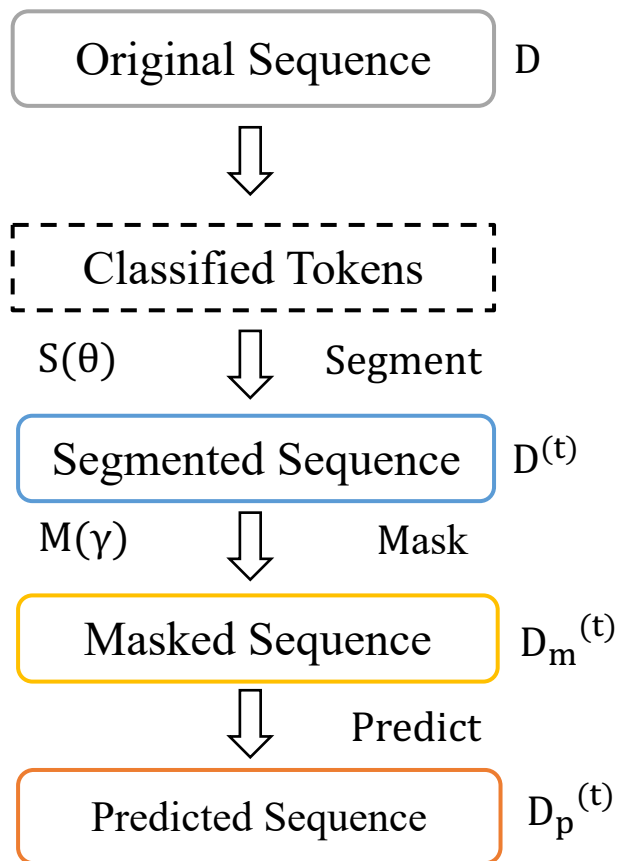


## Model Architecture



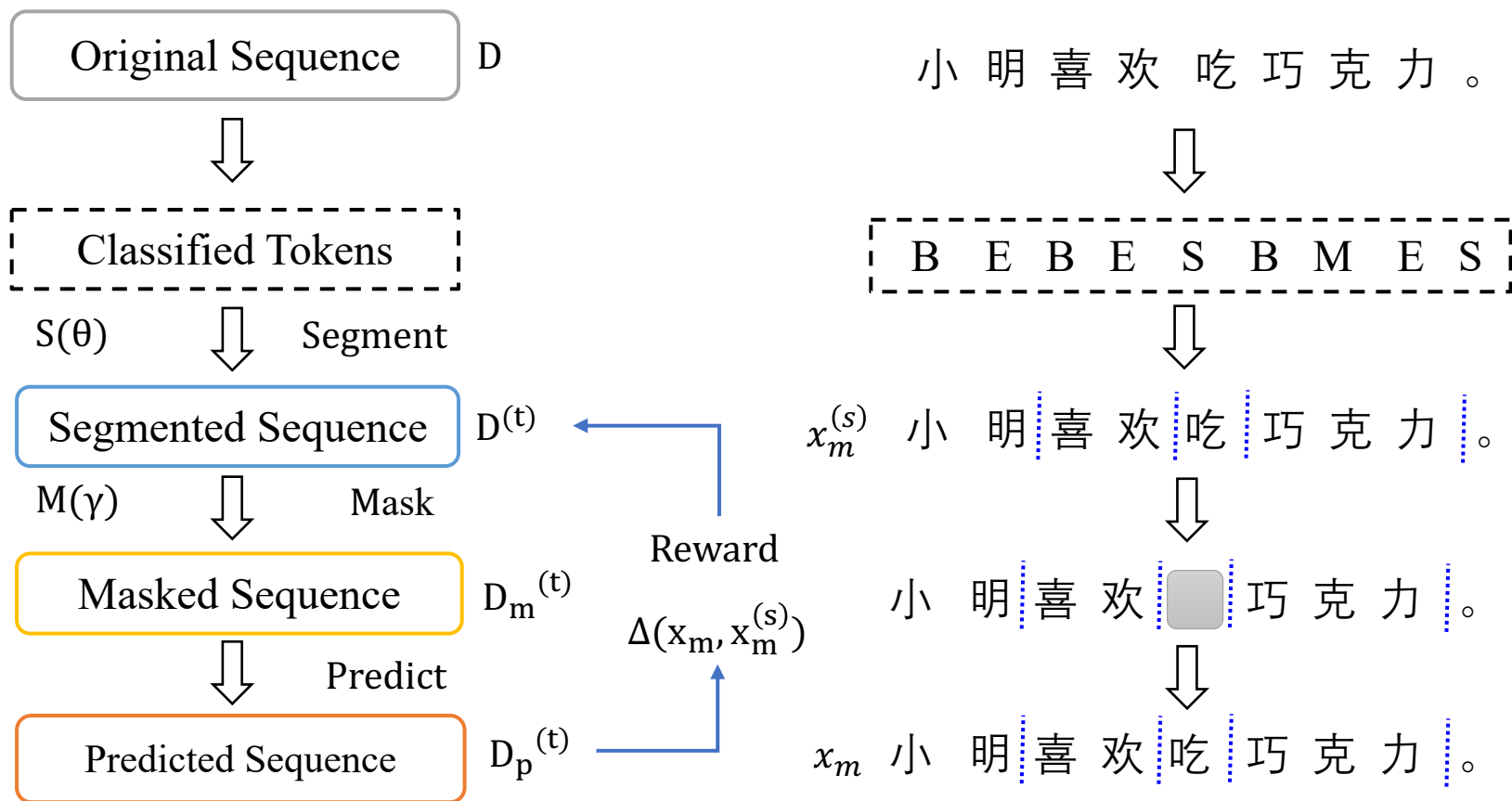


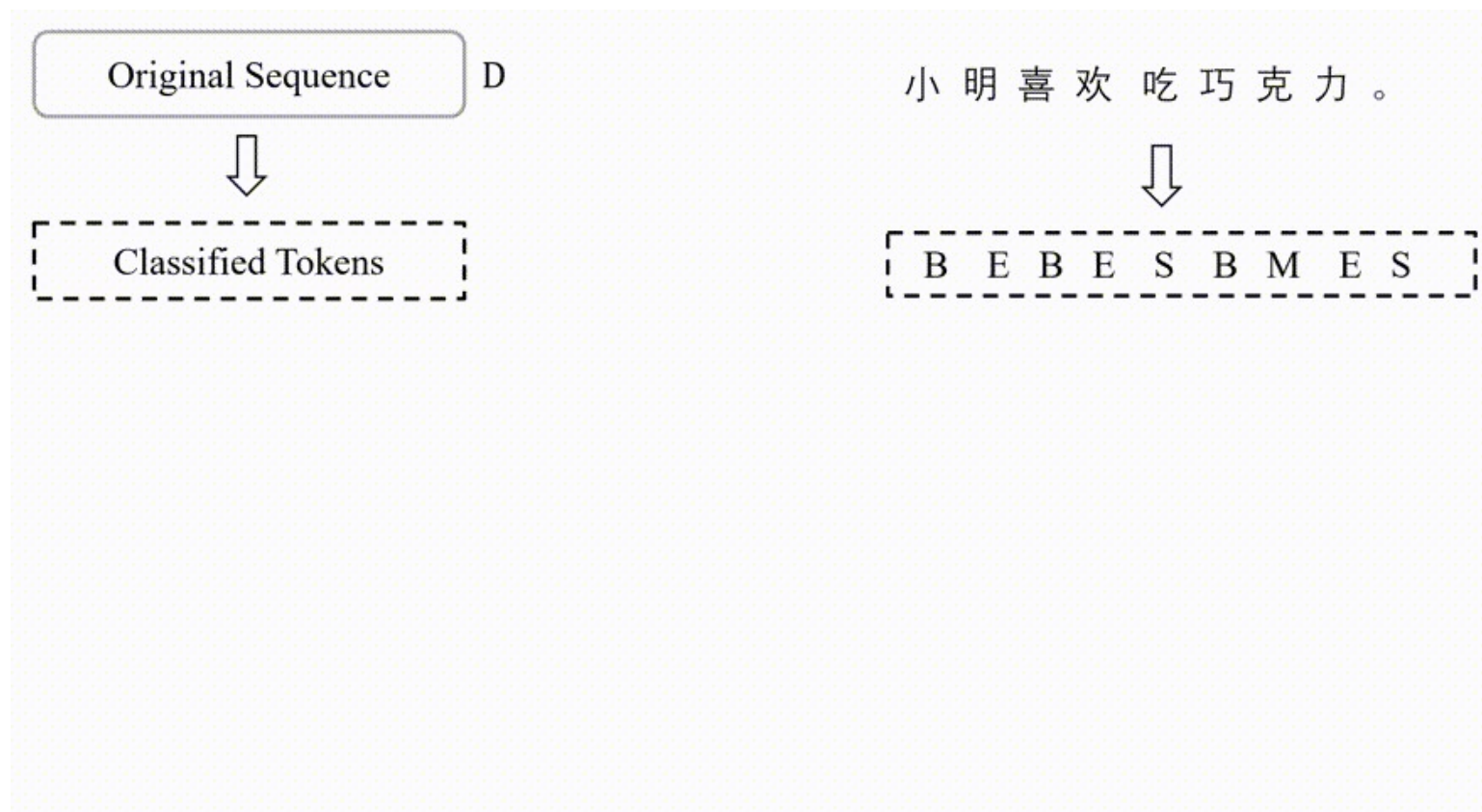
## Model Architecture





## Model Architecture









# Outline

- Chinese Word Segmentation
- Background & Significance
- Challenges & Motivation
- Methodology
- Experiment & Results
- Conclusion & Future Work

# Experiment Settings



## Experiment settings

### Data Characteristics of the Corpus

Corpora	Train	Dev.	Test	Word			Char		
				Type	Token.	Avglen.	Type	Token.	Avglen.
MSRA	84.80K	2.0K	4.0K	90.10K	2.50M	27.24	5.20K	4.01M	46.62
PKU	19.06K	2.0K	1.9K	58.20K	1.21M	57.82	4.70K	1.83M	95.85
AS	0.7M	2.0K	14.4K	0.14M	5.60M	7.7	6.11K	8.37M	11.80
CITYU	53.02K	2.0K	1.5K	70.76K	1.50M	27.45	4.92K	2.40M	45.33
CTB	24.42K	1.9K	2.0K	47.60K	0.80M	27.67	4.44K	1.30M	45.50
SXU	15.62K	1.5K	3.7K	35.92K	0.64M	30.90	4.28K	1.04M	50.50
CNC	0.21M	25.9K	25.9K	0.14M	7.30M	28.19	6.86K	10.08M	43.28
UDC	4.0K	0.5K	0.5K	20.13K	0.12M	24.67	3.60K	0.20M	39.14
ZX	2.37K	0.8K	1.4K	9.14K	0.12M	26.87	2.61K	0.17M	38.05

# Main Results



## Main results

## 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	<u>96.10</u>	<u>95.95</u>	<u>96.10</u>	<u>96.20</u>	<u>96.30</u>
LSTM+CRF	98.10	96.10	96.00	96.80	96.30	<u>96.55</u>	<u>96.61</u>	96.00	<u>96.40</u>
BERT	<u>96.91</u>	<u>95.34</u>	<u>96.47</u>	<u>97.10</u>	<u>97.27</u>	<u>96.40</u>	<u>96.66</u>	<u>97.23</u>	<u>96.49</u>
SELFATT+SOFT	97.60	95.50	95.70	96.40	<u>97.28</u>	<u>96.60</u>	<u>96.88</u>	<u>97.12</u>	<u>96.50</u>
BERT+LTL	<u>97.53</u>	<u>96.23</u>	<u>97.03</u>	<u>97.63</u>	<u>97.34</u>	<u>96.65</u>	<u>96.89</u>	<u>97.51</u>	<u>96.72</u>
Ours	<b>98.12</b>	<b>96.24</b>	<b>97.30</b>	<b>97.83</b>	<b>97.45</b>	<b>96.97</b>	<b>97.25</b>	<b>97.74</b>	<b>96.82</b>



# Main results

## 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	<u>97.22</u>	<u>96.06</u>	<u>97.07</u>	<u>97.39</u>	<u>97.36</u>	<u>96.81</u>	<u>96.71</u>	<u>97.48</u>	<u>96.60</u>
BERT+LTL	<u>96.67</u>	<u>96.30</u>	<u>97.16</u>	<u>97.72</u>	<u>97.38</u>	<u>96.90</u>	<u>97.10</u>	<u>97.61</u>	<u>96.81</u>
Ours	<b>98.19</b>	<b>96.32</b>	<b>97.43</b>	<b>97.80</b>	<b>97.66</b>	<b>97.03</b>	<b>97.34</b>	<b>98.25</b>	<b>97.08</b>



# Main results

## 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	<b>97.93</b>	<b>96.18</b>	<b>97.12</b>	<b>97.68</b>	<b>97.32</b>	<b>96.83</b>	<b>97.12</b>	<b>97.63</b>	<b>96.67</b>



## Main results

### Results on Different Domains

Methods	SIGHAN10		
	Finance	Literature	Medicine
Chen et al. (2015b)	95.20	92.89	92.16
Cai et al. (2017)	95.38	92.90	92.10
Huang et al. (2017)	95.81	94.33	92.26
Zhao et al. (2018)	95.84	93.23	93.73
Zhang et al. (2018)	96.06	94.76	94.18
BERT	<u>95.87</u>	<u>95.57</u>	<u>94.66</u>
BERT+LTL	<u>95.96</u>	<u>95.88</u>	<u>94.87</u>
Ours	<b>95.93</b>	<b>95.96</b>	<b>95.08</b>

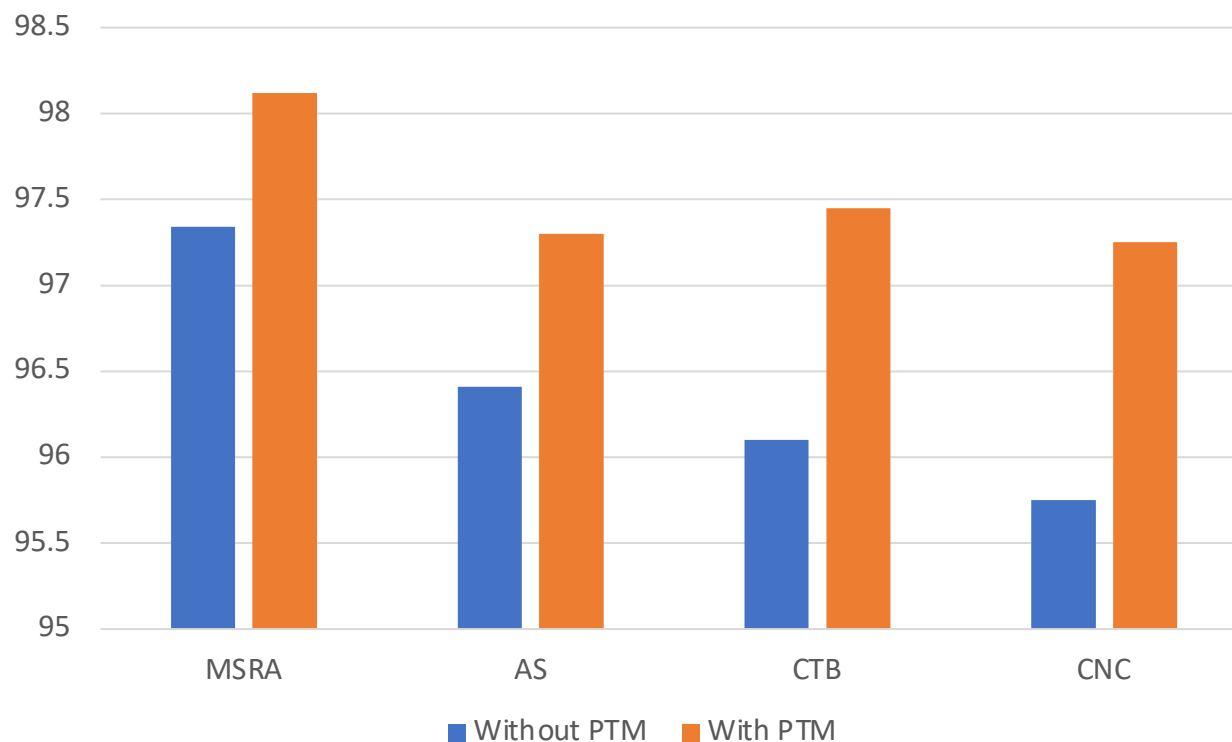


# Ablation Study



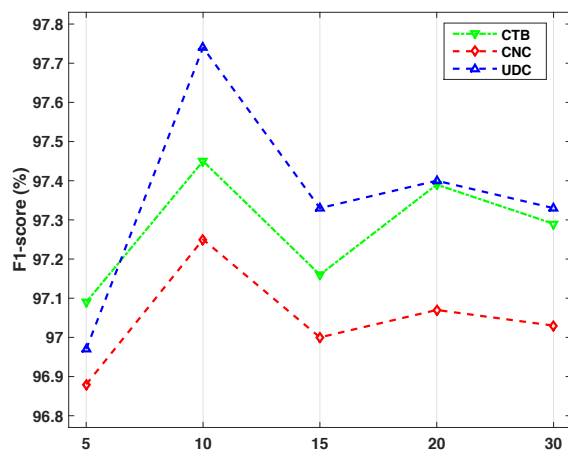
- With and without the PTM

**Effect of Pre-Trained Model**

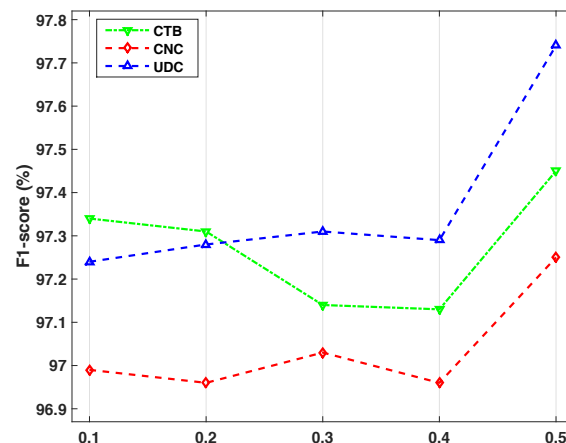




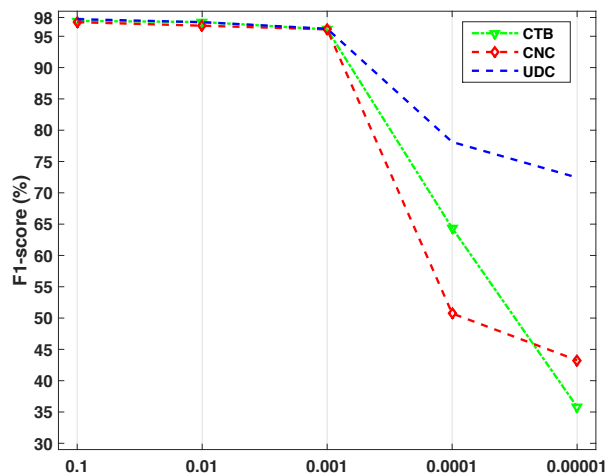
# Effect of hyper-parameters



Size of  $S(x)$



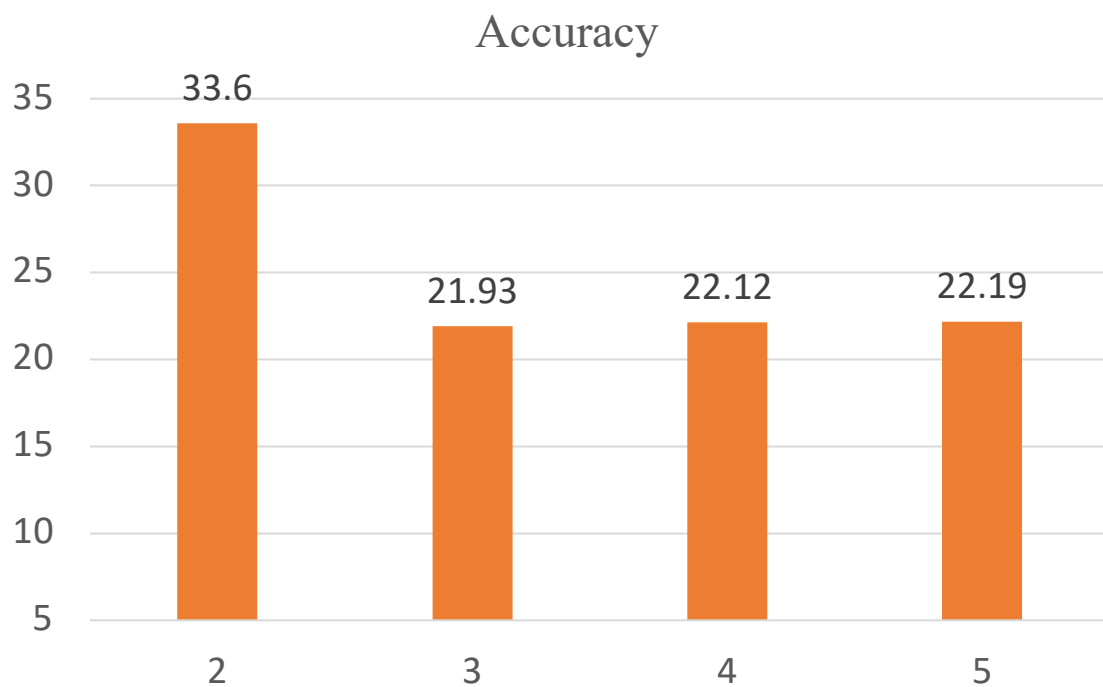
Value of  $\alpha$



Value of  $\lambda$



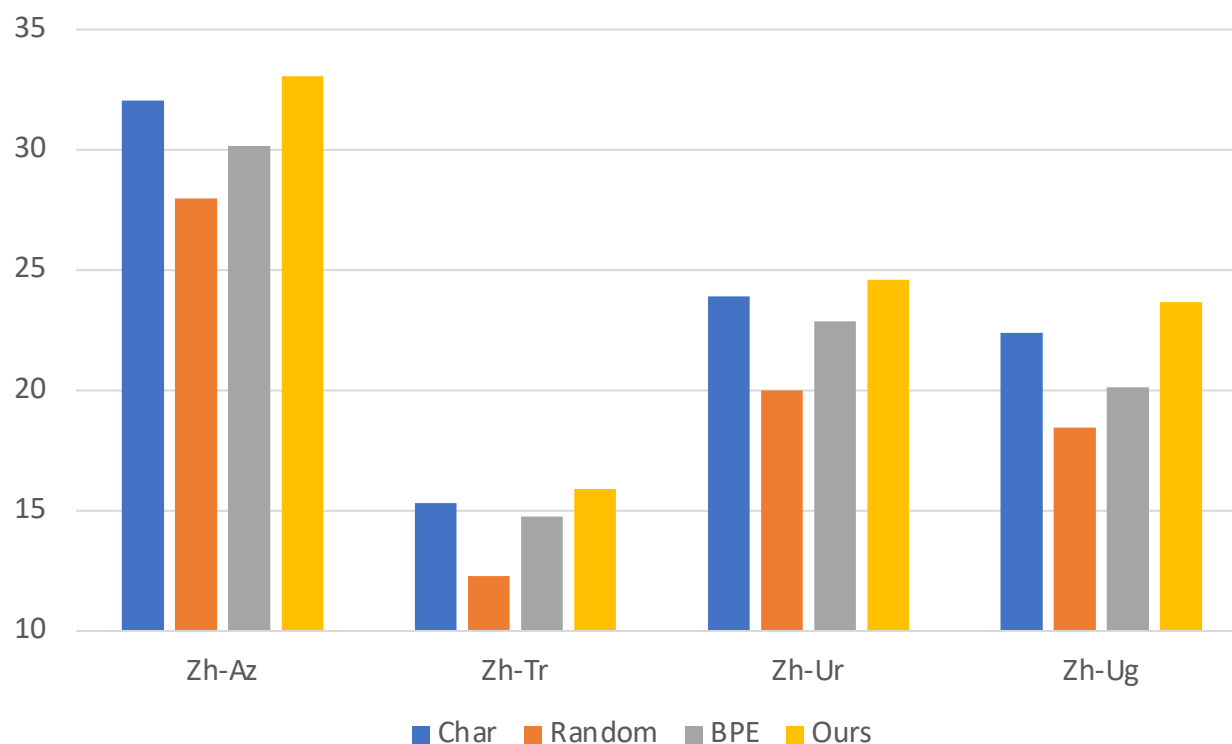
- Effect of masked-count in MLM





## Results on Downstream Task

**Effect of CWS on Low-Resource NMT**





# Outline

- Chinese Word Segmentation
- Background & Significance
- Challenges & Motivation
- Methodology
- Experiment & Results
- Conclusion & Future Work

# Conclusion



- We propose a self-supervised method for CWS, which uses the predictions of revised MLM to assist the word segmentation model.
- We present an improved version of MRT by adding regularization terms to boost the performance of the word segmentation model.
- Experimental results show that our approach outperforms previous methods with different criteria training, and our proposed method also improves the robustness of the model.
- Our method brings positive effects on down stream tasks, such as LRLs NMT.



# Future Work



- In the future, we can also extend our work to the tasks of morphological word segmentation (e.g., morphological analysis).
- It is interesting to make some investigations with the totally unsupervised manner for word segmentation with higher performance.
- We would like to try to design the further steady model by exploiting the lower memory.

## Related Links



Homepage



Paper



Poster



Blog



Code

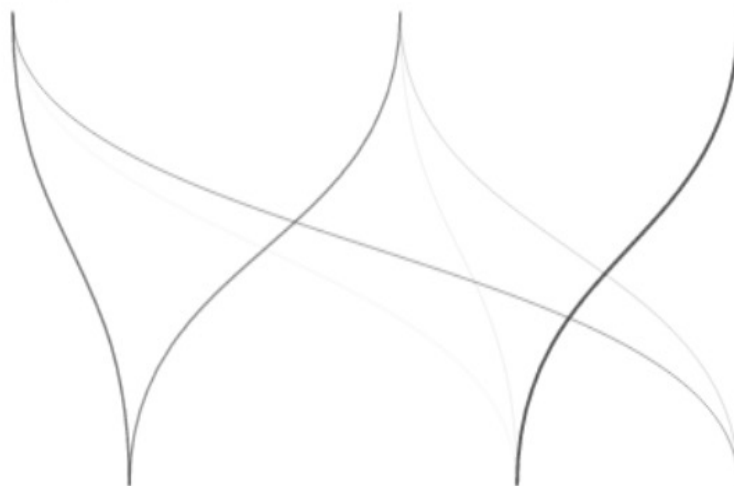


Video

Scan them use WeChat

# Thank You!

Any Questions ?



Questions diversifies ?

This inspiration comes from Dzmitry Bahdanau @ ICLR2014