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

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

(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.

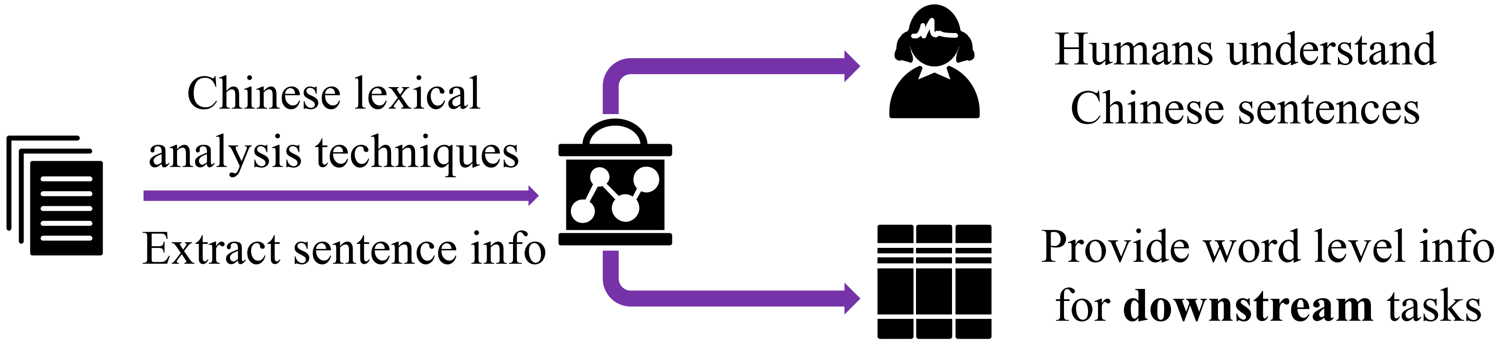
surface form		*underlying form*		*segmentation*
Unquestionably	----	Unquestionablely	-----	un question able ly
				prefix *stem* *suffix* *suffix*

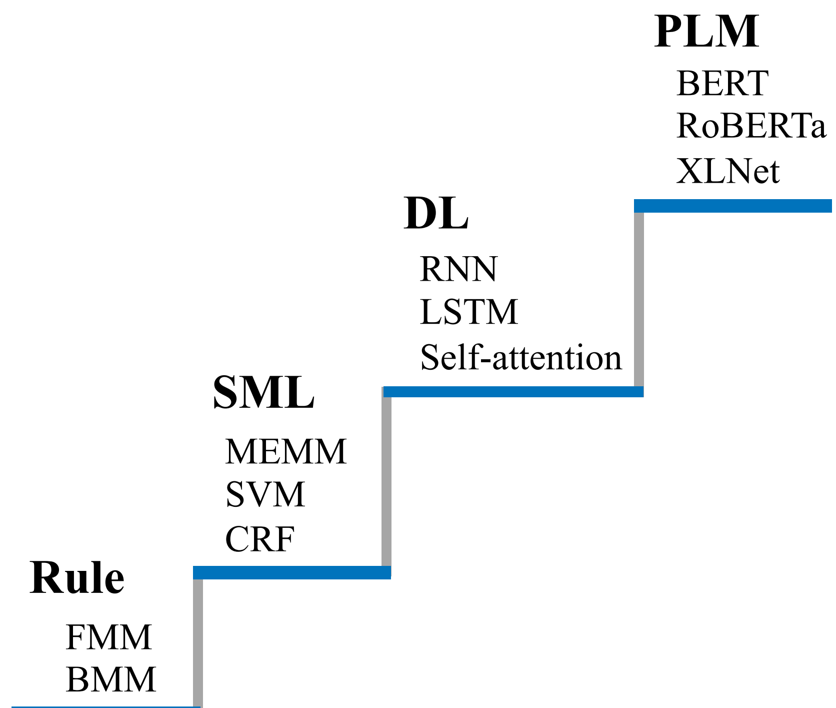
Original		*segmentation*
毫无疑问的	----------	毫无/疑问/的

Outline

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

Background


Chinese lexical analysis techniques
Extract sentence info
Humans understand Chinese sentences
Provide word level info for **downstream** tasks


PLM
BERT
RoBERTa
XLNet
DL
RNN
LSTM
Self-attention
SML
MEMM
SVM
CRF
Rule
FMM
BMM

Significance

Does it make sense?

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

Low-Resource Languages NMT

Cross-Lingual Information Retrieval

Does it make sense?

- Academic value

CWS for NMT

Segmentation Method	BLEU (Zh – En)
CHAR	21.16
TEACHER	23.51
CRF	23.37
CONPRUNE	**23.73**

(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	**91.50**	**88.40**	**87.70**

(Huang et al., 2021)

Outline

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

Challenges

Main challenges

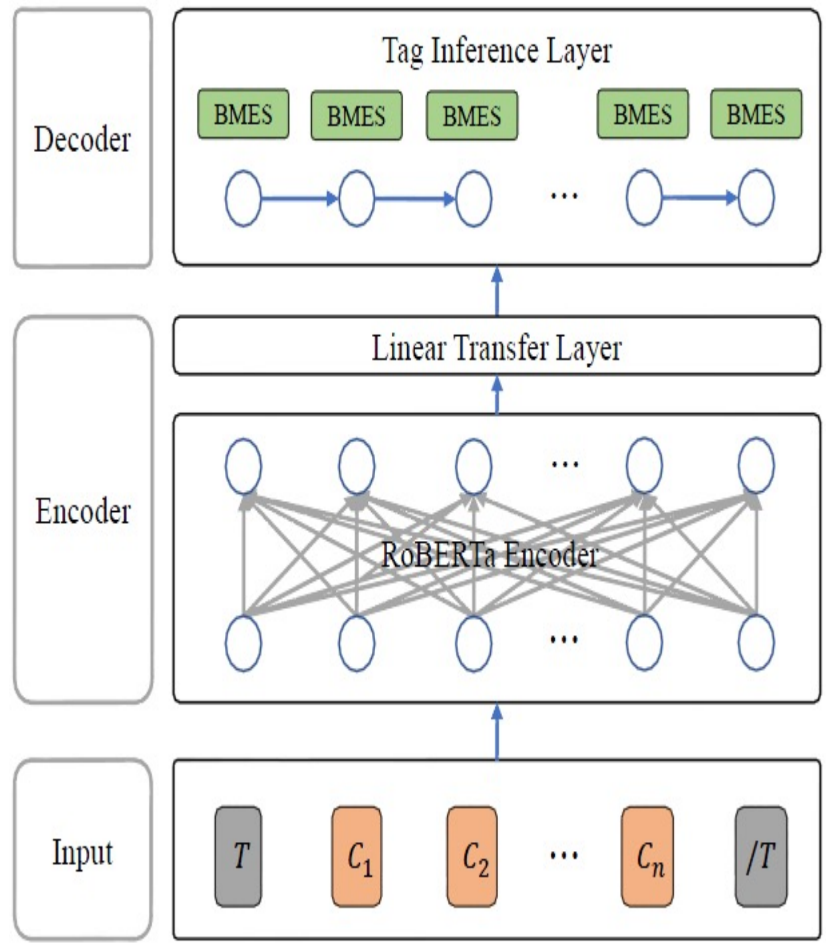
- Annotation inconsistency
  - 操作系统 (operating system) VS. 操作 (operating) /系统 (system)
  - eight times six times
- Word boundary detection
  - 犯罪(crime) / 案(case) 走私案 (smuggling case)

Same sentences in different corpus

<table>
<tr><th>Corpus</th><th>Zhang</th><th>Xiao</th><th>Fan</th><th>attend</th><th colspan="2">a tournament</th></tr>
<tr><td>PKU</td><td>张</td><td colspan="2">小凡</td><td>参加</td><td>比武</td><td>大会</td></tr>
<tr><td>MSRA</td><td colspan="3">张小凡</td><td>参加</td><td colspan="2">比武大会</td></tr>
<tr><td>Zhuxian</td><td colspan="3">张小凡</td><td>参加</td><td>比武</td><td>大会</td></tr>
</table>

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)

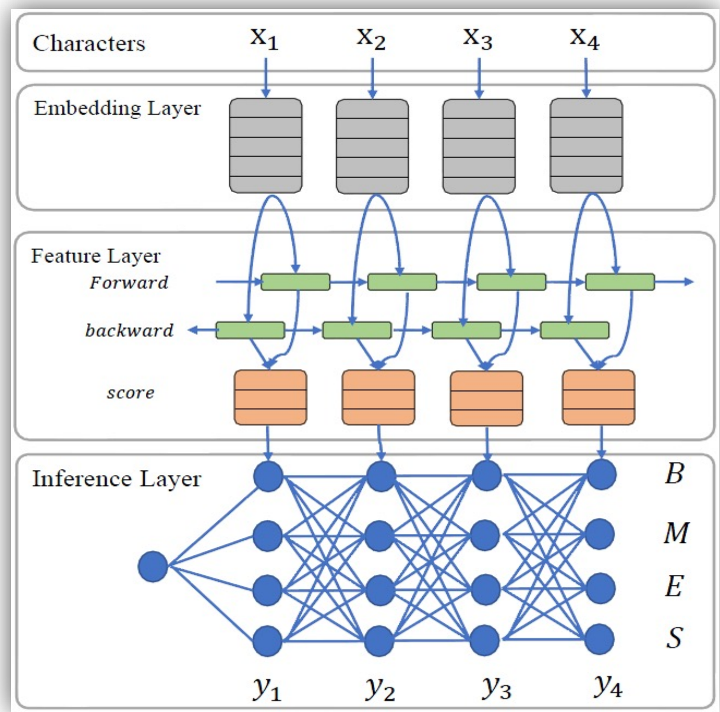
$$X = \{x_1, \cdots, x_n\};\ \ Y^* = \{y_1^*, \cdots, y_n^*\}$$

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

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

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

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

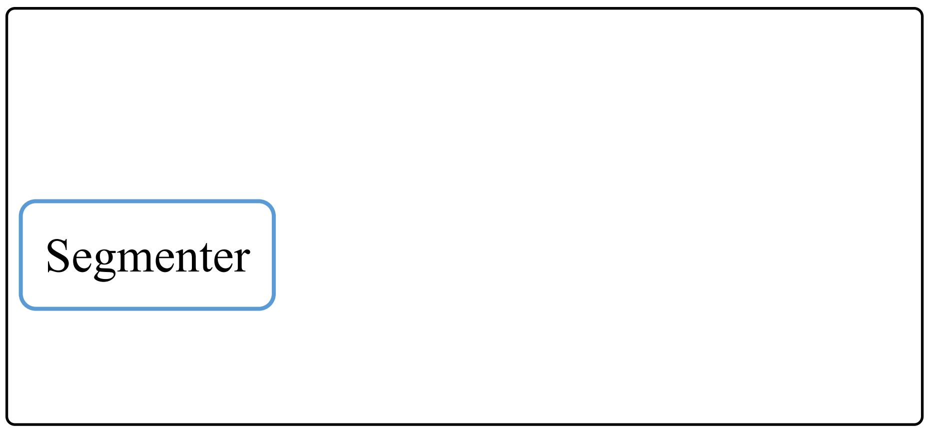


(Chen et al., 2017)

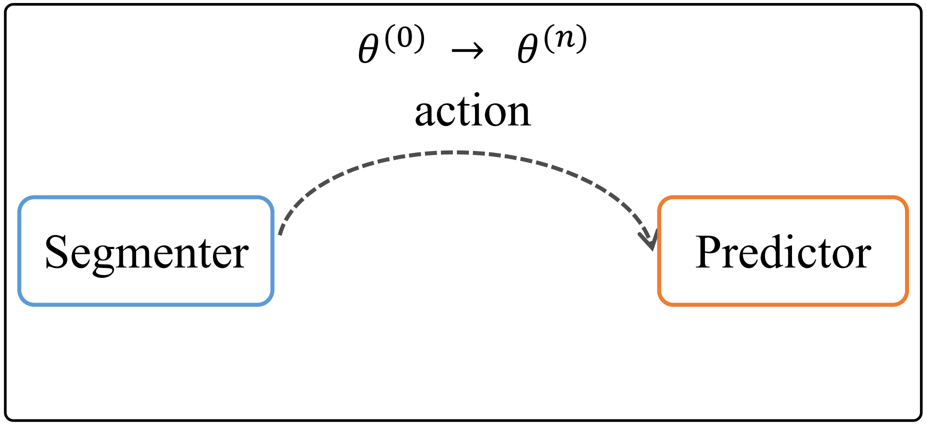
- Output (CRF 4 labels)

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

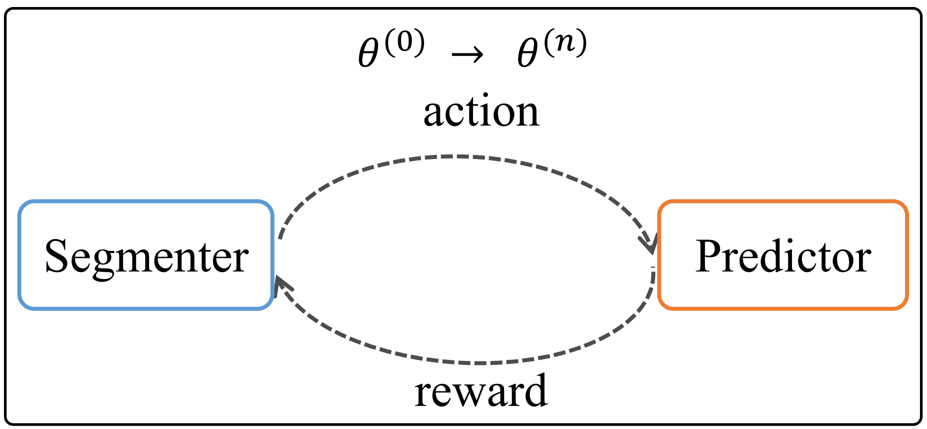
Self-supervised word segmentation model



Self-supervised word segmentation model



Self-supervised word segmentation model




Segmenter

How does it work?

- Input sequence

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

- $\mathbf{x}$ 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.

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

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 D

⇩

Classified Tokens

小 明 喜 欢 吃 巧 克 力 。

⇩

B E B E S B M E S

Model Architecture

Original Sequence D

⇩

Classified Tokens

$S(\theta)$ ⇩ Segment

Segmented Sequence $D^{(t)}$

小 明 喜 欢 吃 巧 克 力 。

⇩

B E B E S B M E S

⇩

$x_m^{(s)}$ 小 明 ⋮ 喜 欢 ⋮ 吃 ⋮ 巧 克 力 ⋮ 。

Model Architecture

Original Sequence D

⇩

Classified Tokens

$S(\theta)$ ⇩ Segment

Segmented Sequence $D^{(t)}$

$M(\gamma)$ ⇩ Mask

Masked Sequence $D_m^{(t)}$

小 明 喜 欢 吃 巧 克 力 。

⇩

B E B E S B M E S

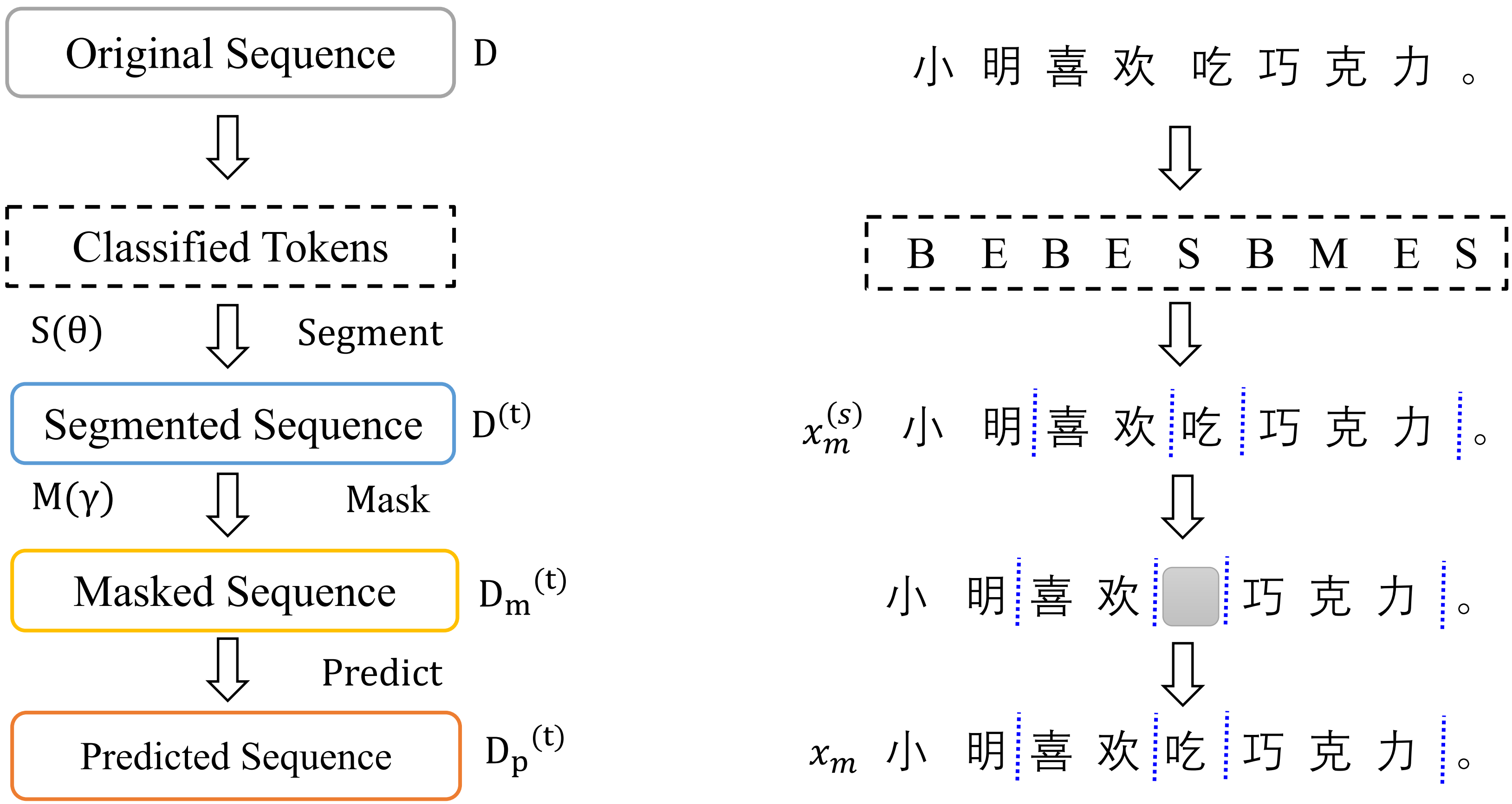
⇩

$x_m^{(s)}$ 小 明 ⁞ 喜 欢 ⁞ 吃 ⁞ 巧 克 力 ⁞ 。

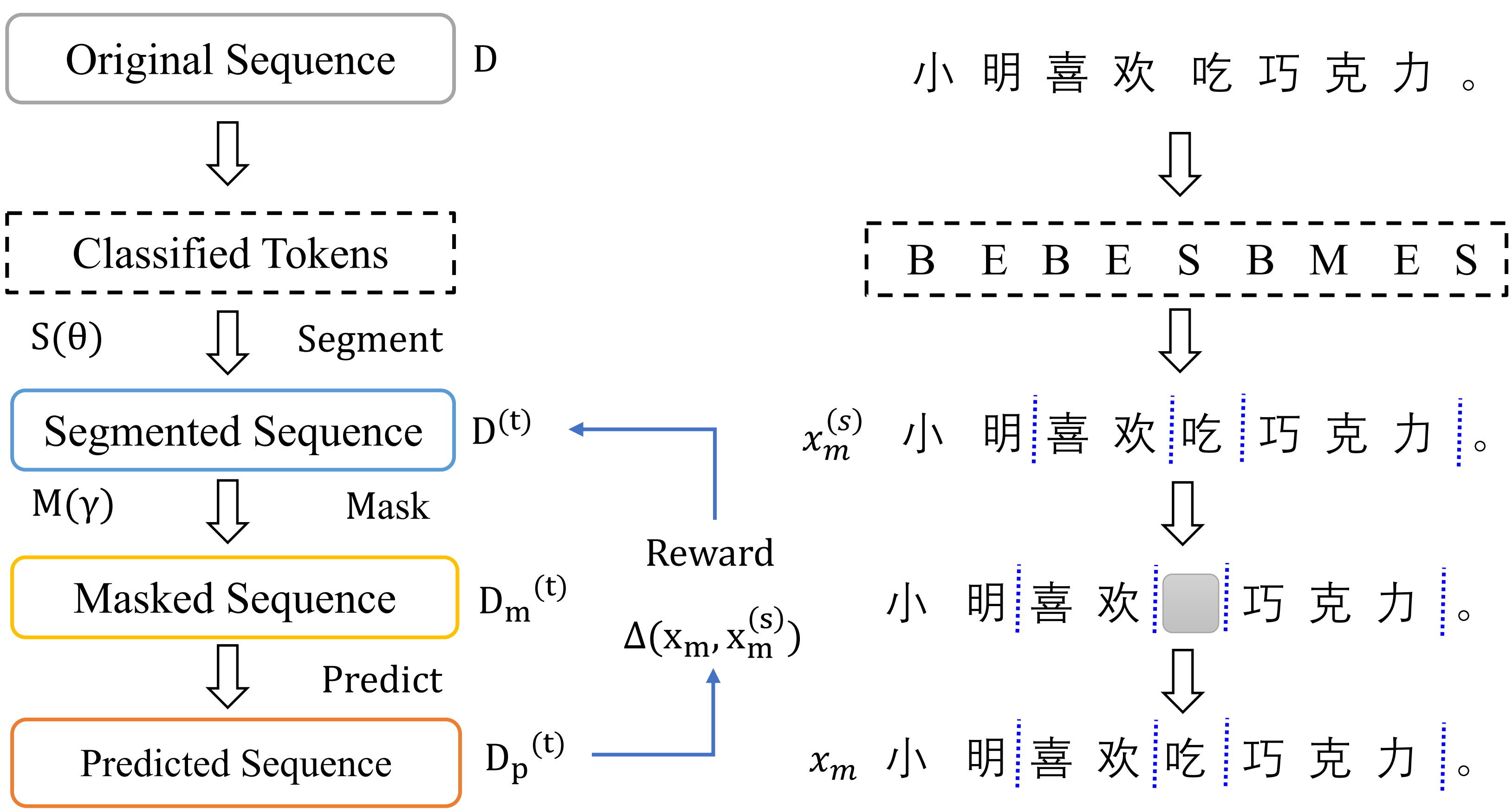
⇩

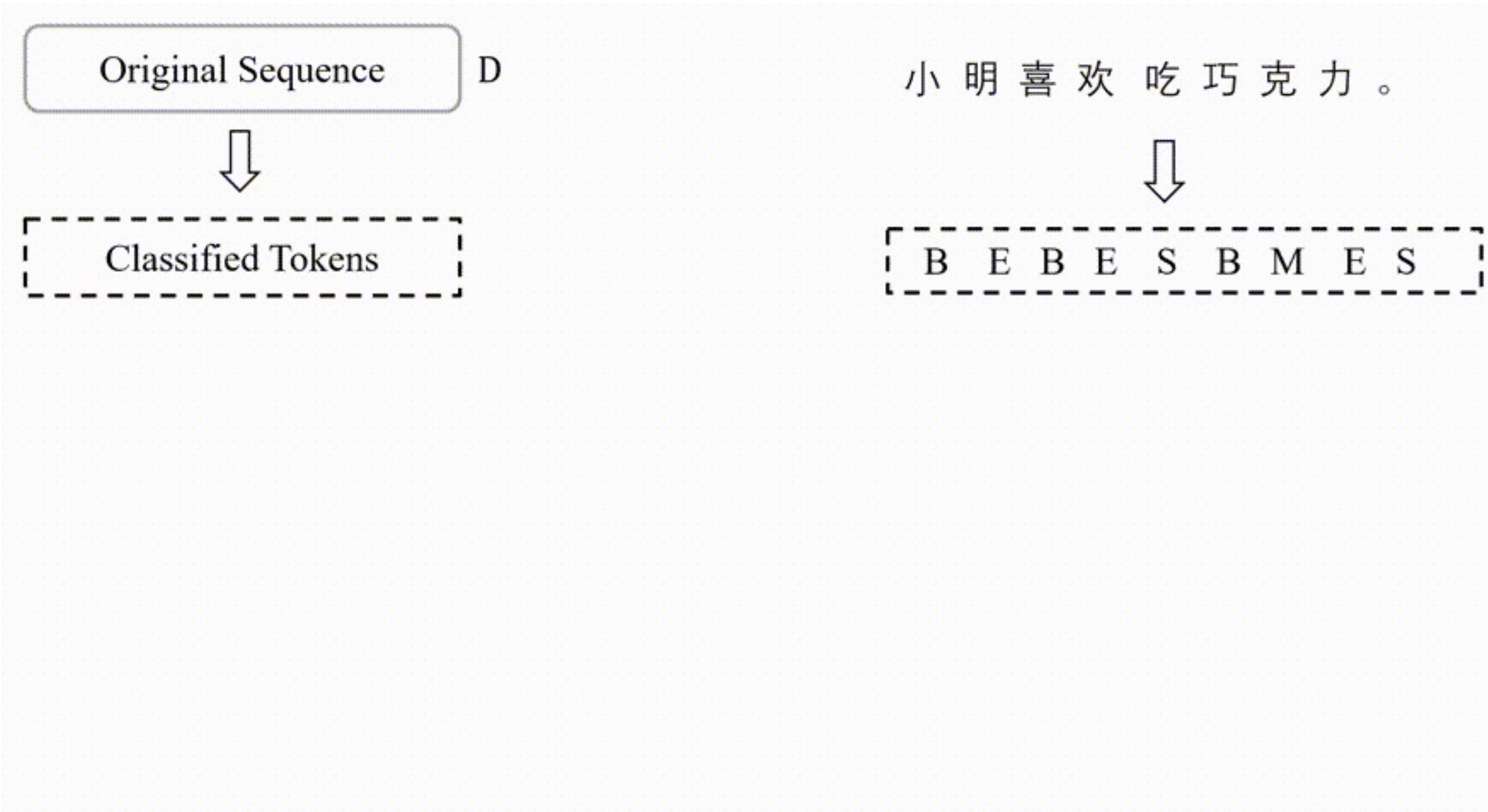
小 明 ⁞ 喜 欢 ⁞ ▢ ⁞ 巧 克 力 ⁞ 。

Model Architecture



Model Architecture




Original Sequence
D
Classified Tokens
小 明 喜 欢 吃 巧 克 力 。
B E B E S B M E S

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	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**

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	**98.19**	**96.32**	**97.43**	**97.80**	**97.66**	**97.03**	**97.34**	**98.25**	**97.08**

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	**97.93**	**96.18**	**97.12**	**97.68**	**97.32**	**96.83**	**97.12**	**97.63**	**96.67**

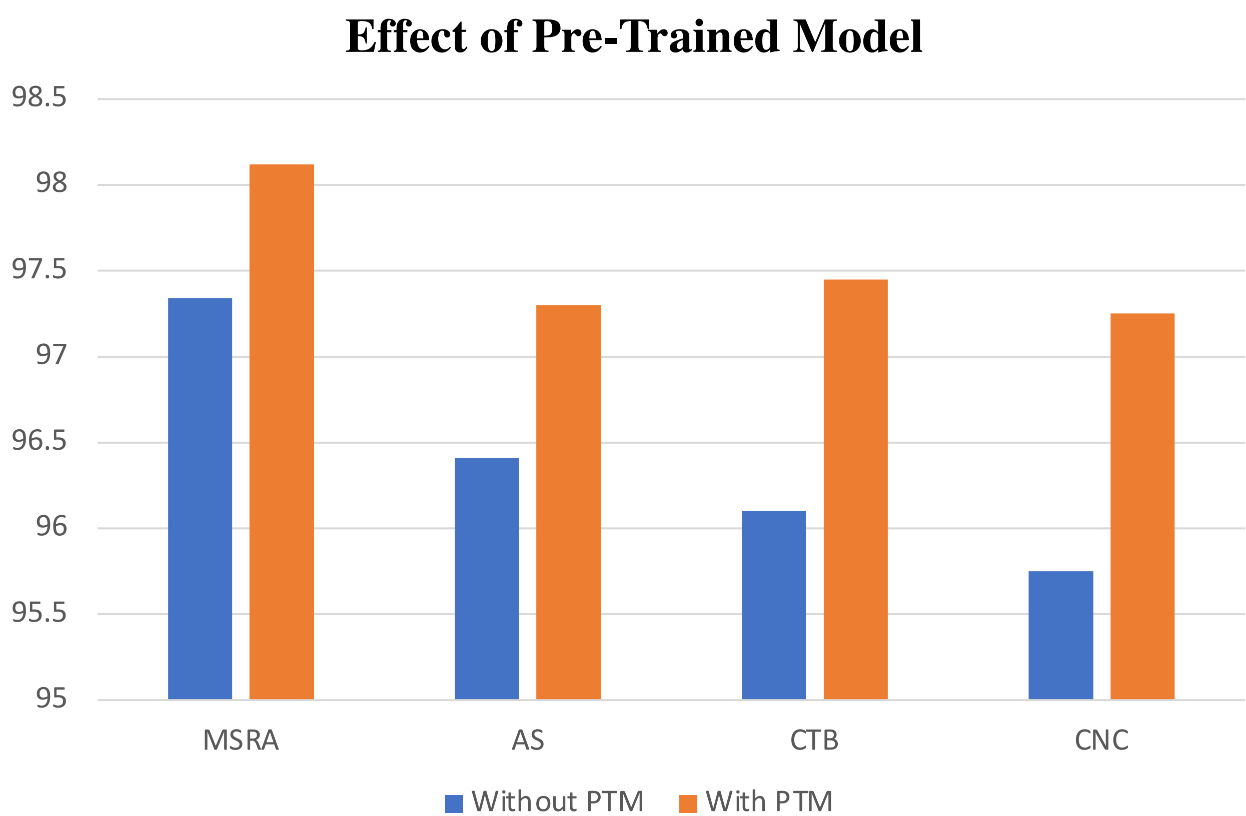
Main results

Results on Different Domains

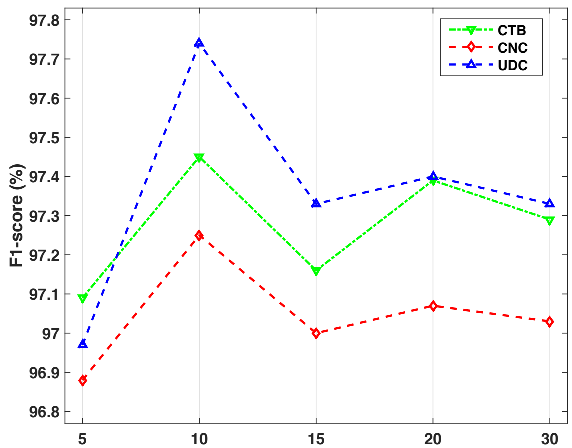
Methods	**SIGHAN10**		
	Finance	Literature	Medicine
Chen et al. (2015b)	95.20	92.89	92.16
Cai et al. (2017)	95.38	92.90	92.10
Huang et al. (2017)	95.81	94.33	92.26
Zhao et al. (2018)	95.84	93.23	93.73
Zhang et al. (2018)	96.06	94.76	94.18
BERT	<u>95.87</u>	<u>95.57</u>	<u>94.66</u>
BERT+LTL	<u>95.96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

Ablation Study

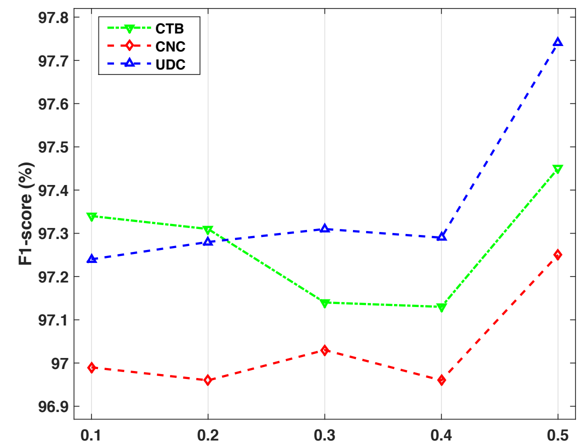
- With and without the PTM



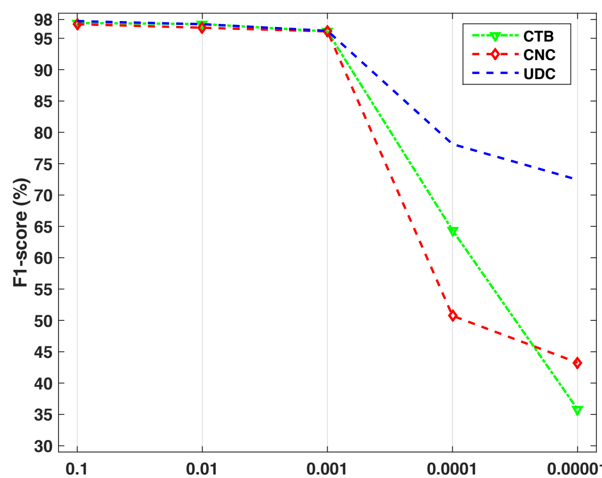
- Effect of hyper-parameters



Size of $S(x)$

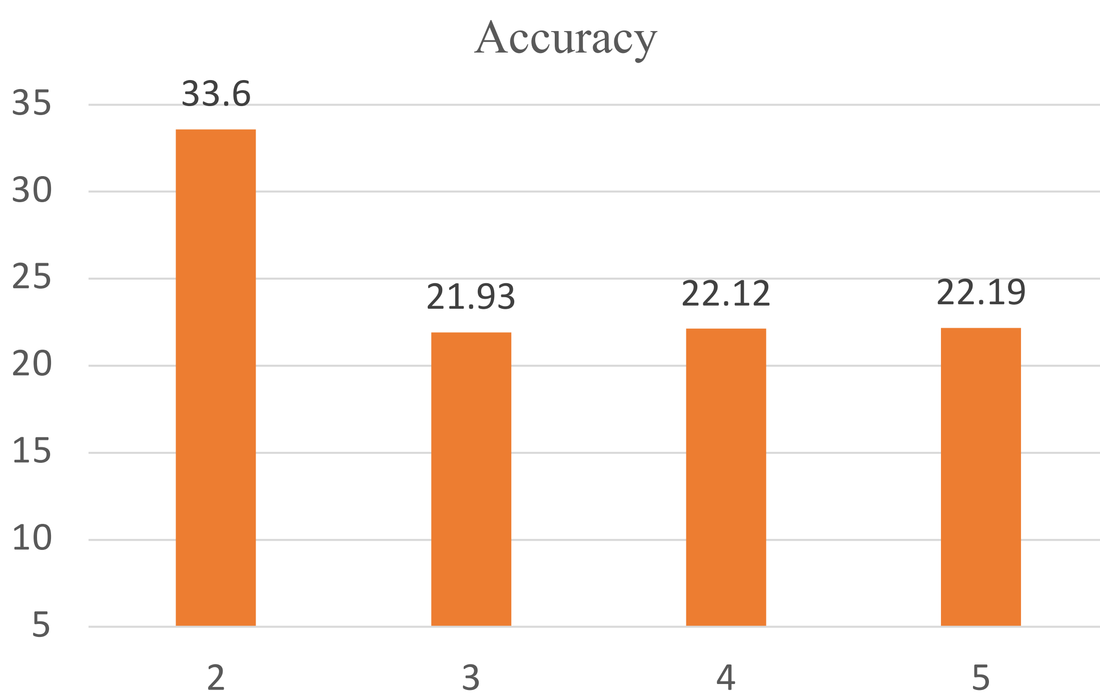


Value of α

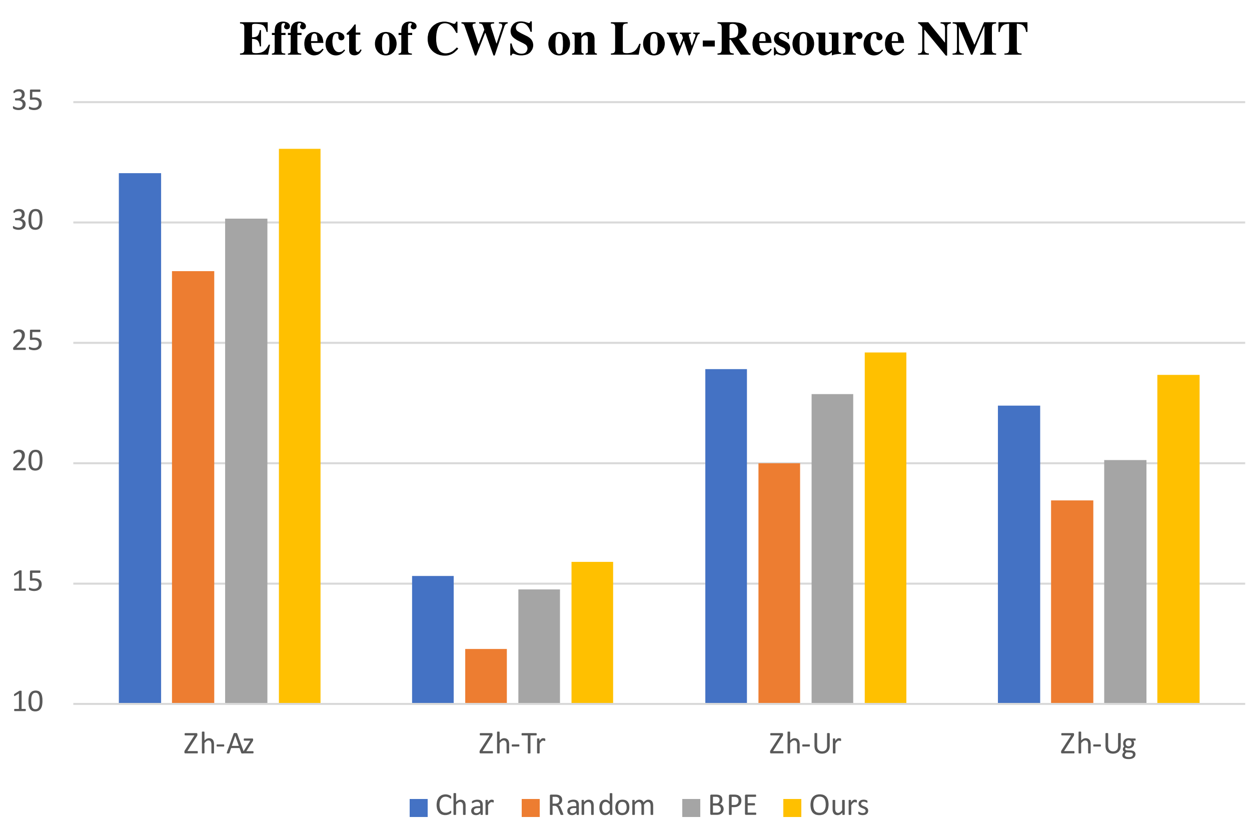


Value of λ

- Effect of masked-count in MLM



Results on Downstream Task



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

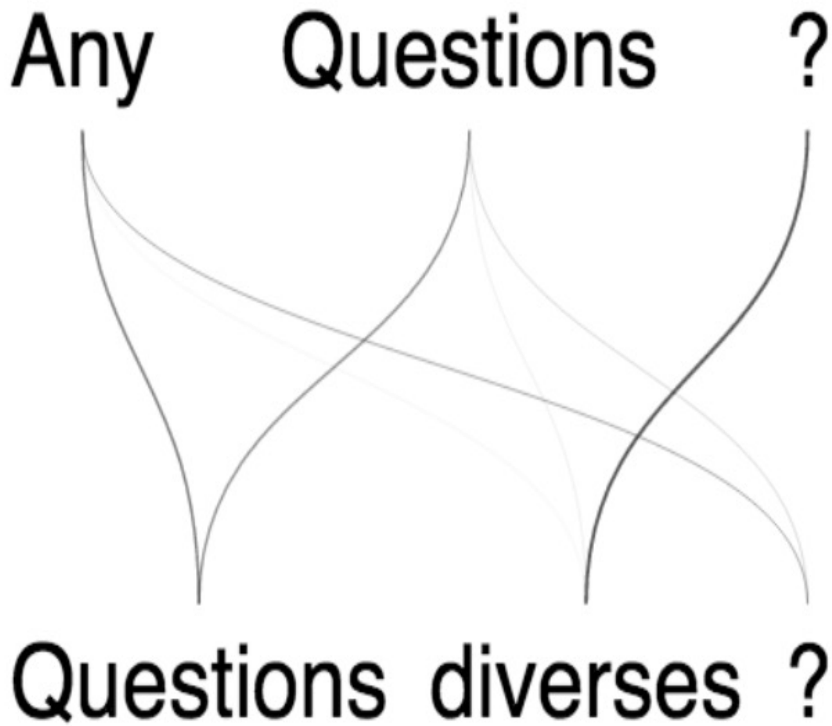
Blog

Code

Video

Scan them use WeChat

Thank You!



This inspiration comes from Dzmitry Bahdanau @ ICLR2014