Masked Conditional Random Fields for Sequence Labeling

Tianwen Wei (魏天闻)　Jianwei Qi (杞坚玮)
Shenghuan He (何声欢)　Songtao Sun (孙松涛)

Xiaomi AI

Table of Contents

Named Entity Recognition

The number of plastic surgeries in Brazil has jumped 30 percent to an estimated 150,000 this year since an anti-inflation plan was introduced in July 1994 , Farid Hakme , the president of the Brazilian Plastic Surgery Society , said.

Named Entities:

- Location: Brazil
- Person: Farid Hakme
- Organization: Brazilian Plastic Surgery Society

Slot Filling

- How to say *give me five* in French?
- I wanna book a flight for tomorrow morning around 9 am, from Wuhan to Beijing.
- Search for Vaccines for Covid-19 in browser

Tagging Scheme

Token	Label	Predict
in	O	O
July	O	O
1994	O	O
,	O	O
Farid	B-PER	B-PER
Hakme	I-PER	I-PER
,	O	O
the	O	O
president	O	O
of	O	O
the	O	O
Brazilian	B-ORG	B-MISC
Plastic	I-ORG	I-ORG
Surgery	I-ORG	I-ORG
Society	I-ORG	I-ORG
(	O	O
SBCP	B-ORG	B-ORG
)	O	O
,	O	O
said	O	O
.	O	O

Figure: Typically a tagging scheme such as BIO or BIOES is used to distinguish the boundary and the type of the text chunk.

Illegal Transition

Token	Label	Predict
in	O	O
July	O	O
1994	O	O
,	O	O
Farid	B-PER	B-PER
Hakme	I-PER	I-PER
,	O	O
the	O	O
president	O	O
of	O	O
the	O	O
Brazilian	B-ORG	B-MISC
Plastic	I-ORG	I-ORG
Surgery	I-ORG	I-ORG
Society	I-ORG	I-ORG
(	O	O
SBCP	B-ORG	B-ORG
)	O	O
,	O	O
said	O	O
.	O	O

Figure: In this example, the model prediction contains an illegal transition B-MISC $\rightarrow$ I-ORG.

Path

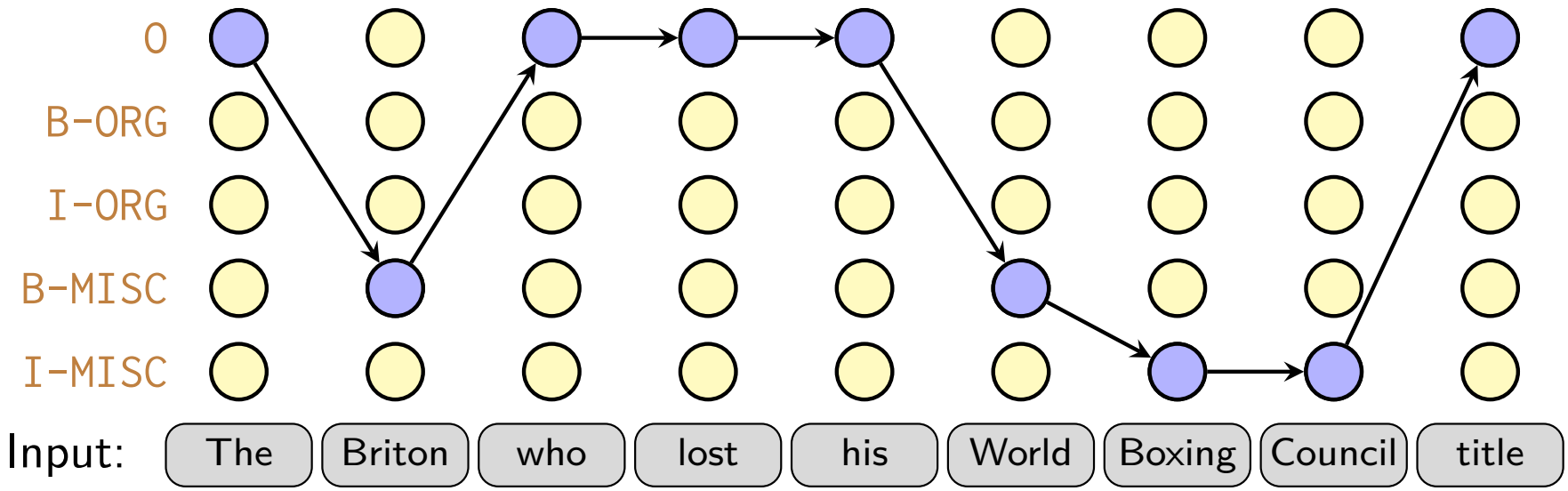


Figure: A label sequence $(y_1, \ldots, y_T)$ can be regarded as a **path**. If a path contains at least one illegal transition, then we say it is an **illegal path**.

Default Approach

Sang et al. (2000) has noted that

> *"The output of a chunk recognizer may contain inconsistencies in the chunk tags in case a word tagged* `I-X` *follows a word tagged* `O` *or* `I-Y`*, with* `X` *and* `Y` *being different. These inconsistencies can be resolved by assuming that such* `I-X` *tags starts a new chunk."*

Before:	O	I-PER	O	B-LOC	I-MISC
After:	O	B-PER	O	B-LOC	B-MISC

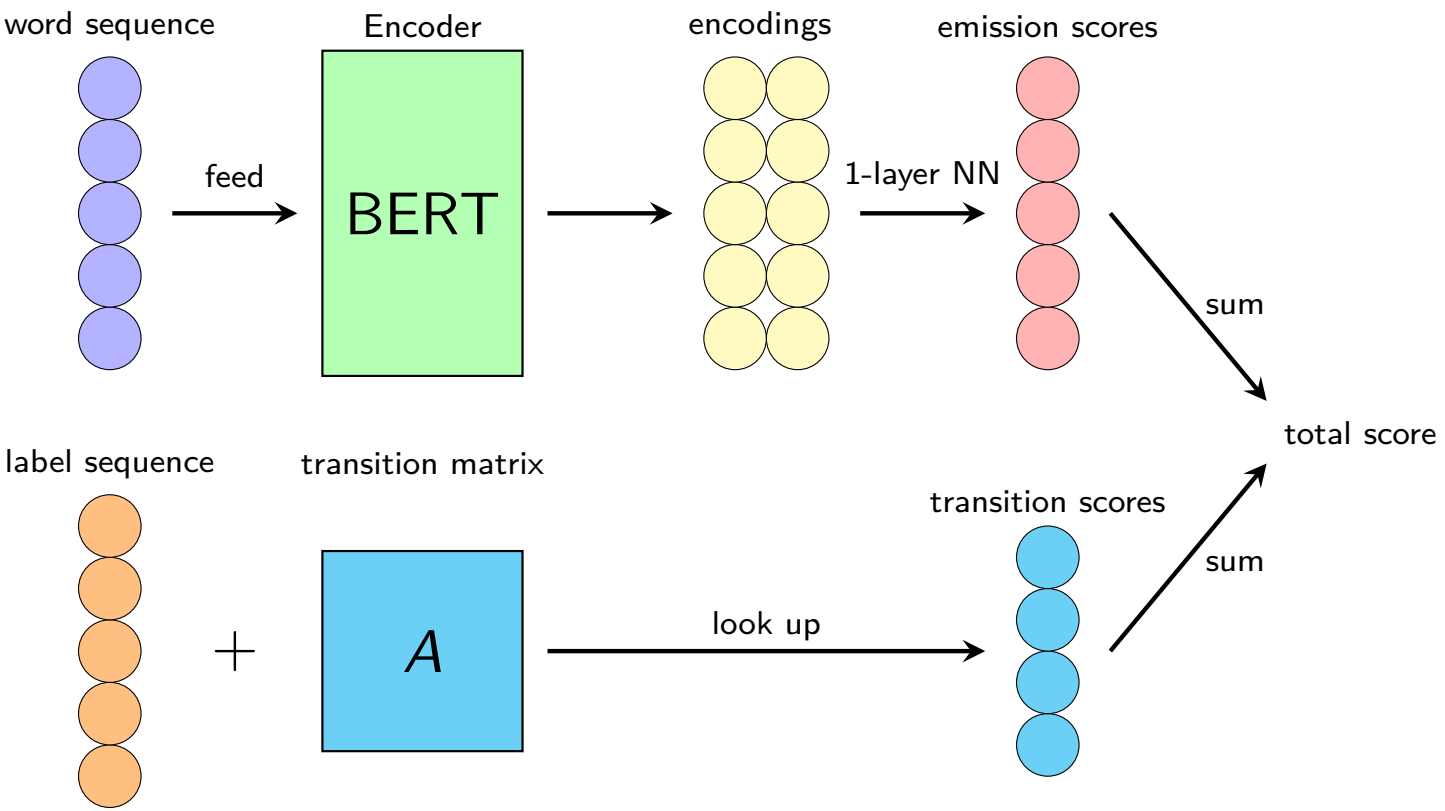
Statistics on Illegal Paths

Dataset	legal & TP	illegal & TP	legal & FP	illegal & FP	$\frac{\text{illegal \& TP}}{\text{illegal}}$	$\frac{\text{illegal \& FP}}{\text{FP}}$	$\frac{\text{illegal}}{\text{total}}$
Resume	1445	1	68	17	1.4%	20%	1.2%
MSRA	5853	6	318	107	1.9%	25%	1.8%
Ontonotes	5323	5	1336	314	1.6%	19%	4.6%
Weibo	277	2	124	46	1.6%	27%	10.7%
ATIS	1643	0	70	24	0.0%	26%	1.4%
SNIPS	1542	13	237	156	5.2%	40%	8.7%
CoNLL2000	22957	36	888	100	3.9%	10%	0.6%
CoNLL2003	5131	2	535	74	0.4%	12%	1.3%

Figure: Up to 40% of false positives are due to illegal paths.

Table of Contents

Conventional Neural-CRF Models



The CRF model assigns a score $s(y, x)$ for any input sequence x and any label sequence y. The loss of a labeled sample (x, y) is defined as:

$$\mathcal{L}_{(x,y)} = -\log \frac{\exp s(y, x)}{\sum_{p \in \mathcal{P}} \exp s(p, x)} \tag{2.1}$$

CRF Training and Decoding

Denote by A the transition matrix and by W the collection of all remaining trainable parameters in the neural network.

- **Training**

$$\mathcal{L}(W, A) = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \log \frac{\exp s(y, x)}{\sum_{p\in\mathcal{P}} \exp s(p, x)} \tag{2.2}$$

- **Decoding**

$$y_{\text{opt}} = \underset{p\in\mathcal{P}}{\operatorname{argmax}}\ s(p, x_{\text{test}}, W_{\text{opt}}, A_{\text{opt}}). \tag{2.3}$$

Table of Contents

Our Approach

Let $\mathcal{I}$ denote the set of illegal paths, i.e. paths that contain at least one illegal transition. We propose to constrain the "space of candidate paths" to be the set of all legal paths $\mathcal{P}/\mathcal{I}$:

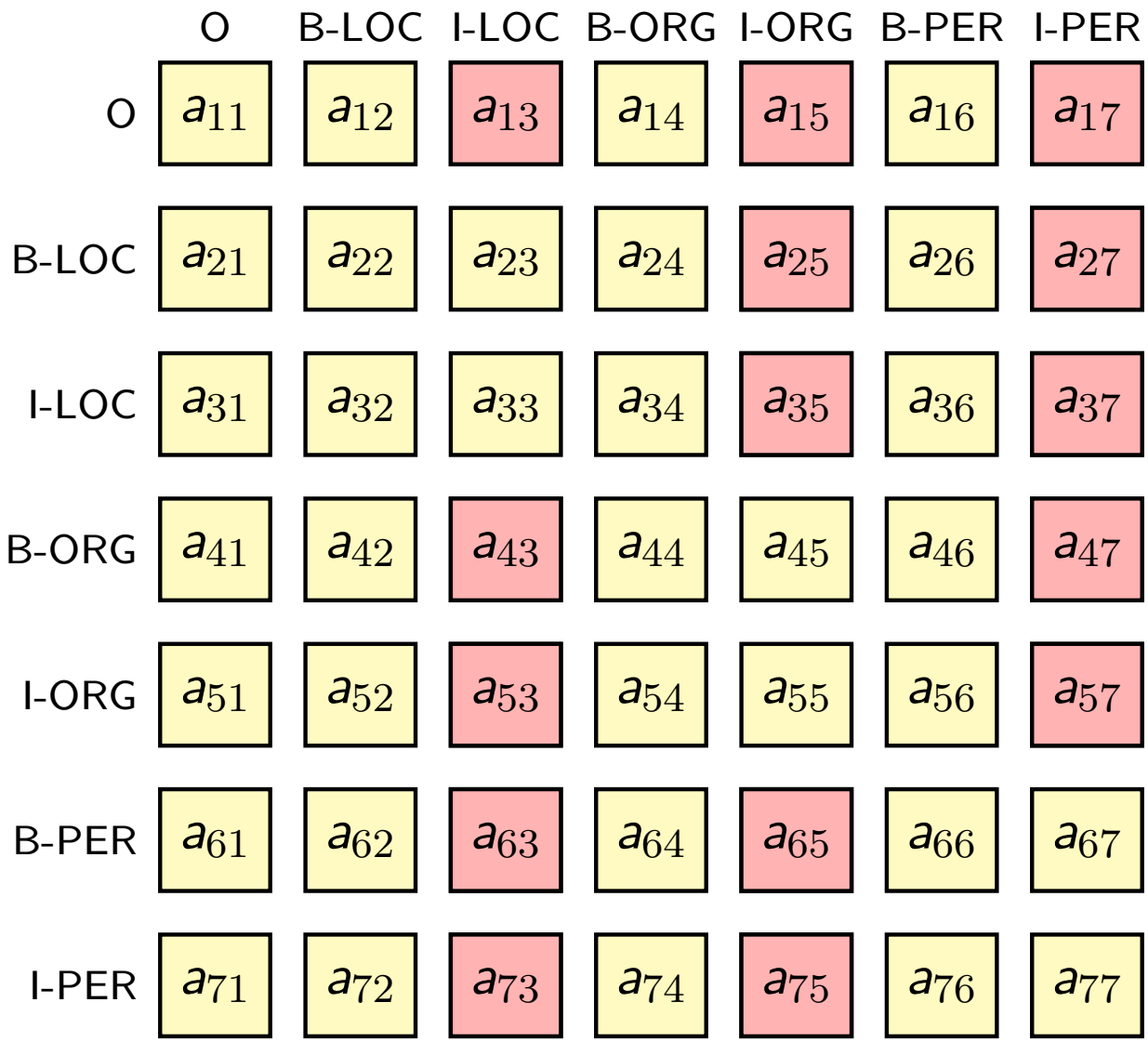
- **Training**

$$\mathcal{L}'(W, A) = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \log \frac{\exp s(y, x)}{\sum_{p\in\mathcal{P}/\mathcal{I}} \exp s(p, x)} \tag{3.4}$$

- **Decoding**

$$y_{\text{opt}} = \underset{p\in\mathcal{P}/\mathcal{I}}{\text{argmax}}\ s(p, x_{\text{test}}, W_{\text{opt}}, A_{\text{opt}}). \tag{3.5}$$

Masked Transition Matrix



Main Result

Proposition

Denote by $\Omega \subset [d] \times [d]$ *the set of all illegal transitions. For a given transition matrix* A*, we denote by* $\bar{A}(c) = \left(\bar{a}_{ij}(c)\right)$ *the* masked transition matrix *of* A *defined as*

$$\bar{a}_{ij}(c) = \begin{cases} c & if\ (i,j) \in \Omega, \\ a_{ij} & otherwise, \end{cases} \tag{3.6}$$

where $c \ll 0$ *is the* transition mask*. Then for arbitrary model weights* (W_0, A_0)*, we have*

$$\lim_{c \to -\infty} \mathcal{L}(W_0, \bar{A}_0(c)) = \mathcal{L}'(W_0, A_0) \tag{3.7}$$

$$\lim_{c \to -\infty} \nabla_W \mathcal{L}(W_0, \bar{A}_0(c)) = \nabla_W \mathcal{L}'(W_0, A_0) \tag{3.8}$$

and for all $(i,j) \in \Omega$

$$\lim_{c \to -\infty} \nabla_{a_{ij}} \mathcal{L}(W_0, \bar{A}_0(c)) = \nabla_{a_{ij}} \mathcal{L}'(W_0, A_0). \tag{3.9}$$

Intuition

This results states that optimizing

$$\mathcal{L}'(W, A) = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \log \frac{\exp s(y, x)}{\sum_{p\in\mathcal{P}/\mathcal{I}} \exp s(p, x)}$$

is (almost) equivalent to optimizing

$$\mathcal{L}(W, A) = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \log \frac{\exp s(y, x)}{\sum_{p\in\mathcal{P}} \exp s(p, x)}$$

if we keep the transition matrix A masked during the optimization process.

MCRF Algorithm

Input: Library for computing the gradients of conventional CRF loss (3.4), training dataset $\mathcal{S}$, stopping criterion $\mathcal{C}$, set of illegal transitions Ω, masking constant $c \ll 0$.
Initialize: model weight W and tag transition matrix $A = (a_{ij})$.
while $\mathcal{C}$ is not met **do**
    Sample a mini-batch from $\mathcal{S}$
    Update W and A based on batch gradient
    for $(i, j) \in \Omega$ **do**
        $a_{ij} \leftarrow c$ ▷ maintain the mask
    end for
end while
Output: Optimized W and A.

CRF vs. MCRF

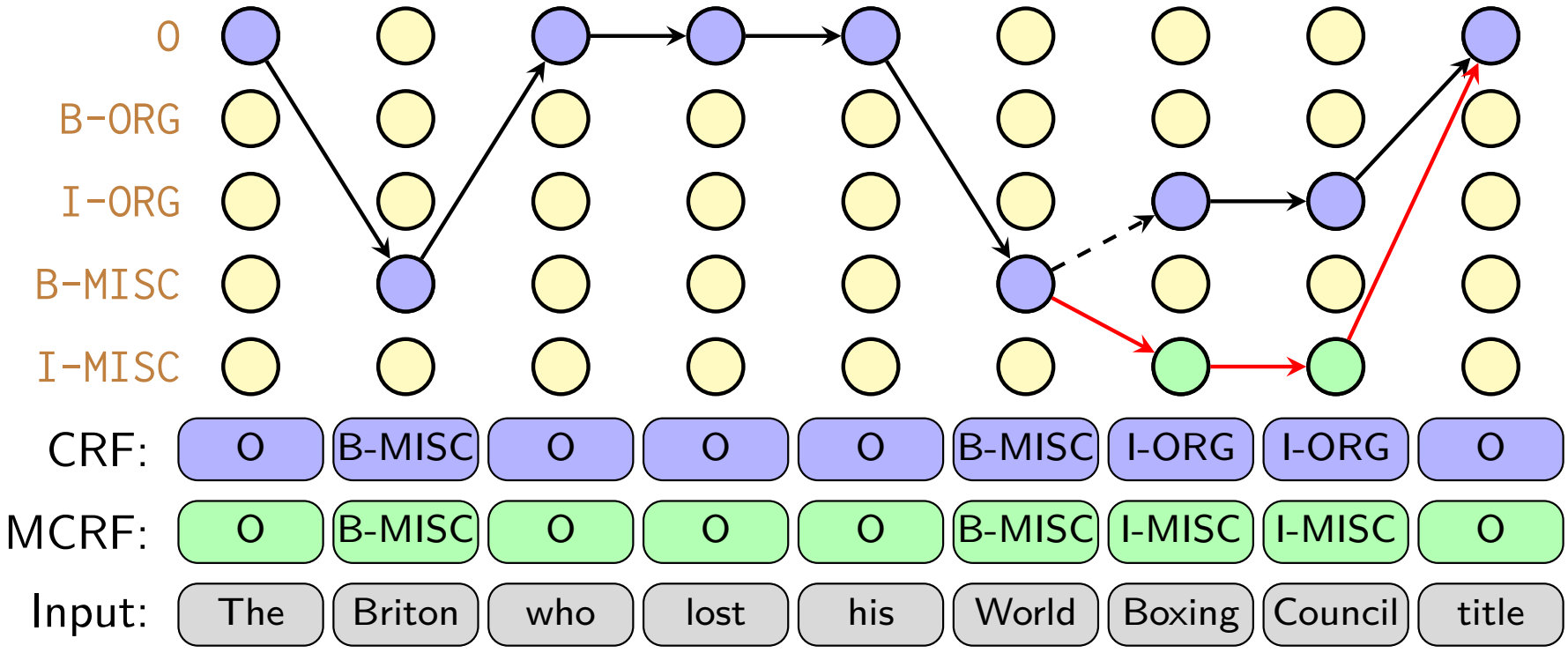


Figure: An example of CRF decoded path vs. MCRF decoded path.

Empirical Experiments

dataset	task	lan.	labels	train	dev	test
Resume	NER	CN	8	3.8k	472	477
MSRA	NER	CN	3	46.3k	-	4.3k
Ontonotes	NER	CN	4	15.7k	4.3k	4.3k
Weibo	NER	CN	7	1.3k	270	270
ATIS	SF	EN	79	4.5k	500	893
SNIPS	SF	EN	39	13.0k	700	700
CoNLL2000	Chunk.	EN	11	8.9k	-	2.0k
CoNLL2003	NER	EN	4	14.0k	3.2k	3.5k

Baselines

- **-tagger** Using softmax as the output layer.
- **-CRF** Using CRF as the output layer.
- **-retain** Keep and retag the illegal segments when encountered.
- **-discard** Discard the illegal segments when encountered.
- **MCRF-decoding** Train as in CRF but apply transition masking in decoding*.
- **MCRF-training** Maintain the transition masking in training.

*Implemented in AllenNLP.

Results on Chinese NER

	Resume	MSRA	Ontonotes	Weibo
Lattice	94.5	93.2	73.9	58.8
Glyce	<u>96.5</u>	95.5	81.6	67.6
SoftLexicon	96.1	95.4	82.8	<u>70.5</u>
FLAT	95.9	96.1	81.8	68.6
MRC	-	95.7	82.1	-
DSC	-	<u>96.7</u>	<u>84.5</u>	-
BERT-tagger-retain	95.7 (94.7)	94.0 (92.7)	78.1 (76.8)	67.7 (65.3)
BERT-tagger-discard	96.2 (95.5)	94.6 (93.6)	80.7 (79.2)	69.7 (67.5)
BERT-CRF-retain	95.9 (94.8)	94.2 (93.7)	81.8 (81.2)	70.8 (64.5)
BERT-CRF-discard	97.2 (96.6)	95.5 (94.9)	83.1 (82.4)	71.9 (65.7)
BERT-MCRF-decoding	97.3 (96.6)	95.6 (95.0)	83.2 (82.5)	72.2 (65.8)
BERT-MCRF-training	**97.6** (96.9)	**95.9** (95.3)	**83.7** (82.7)	**72.4** (66.5)

Loss curve

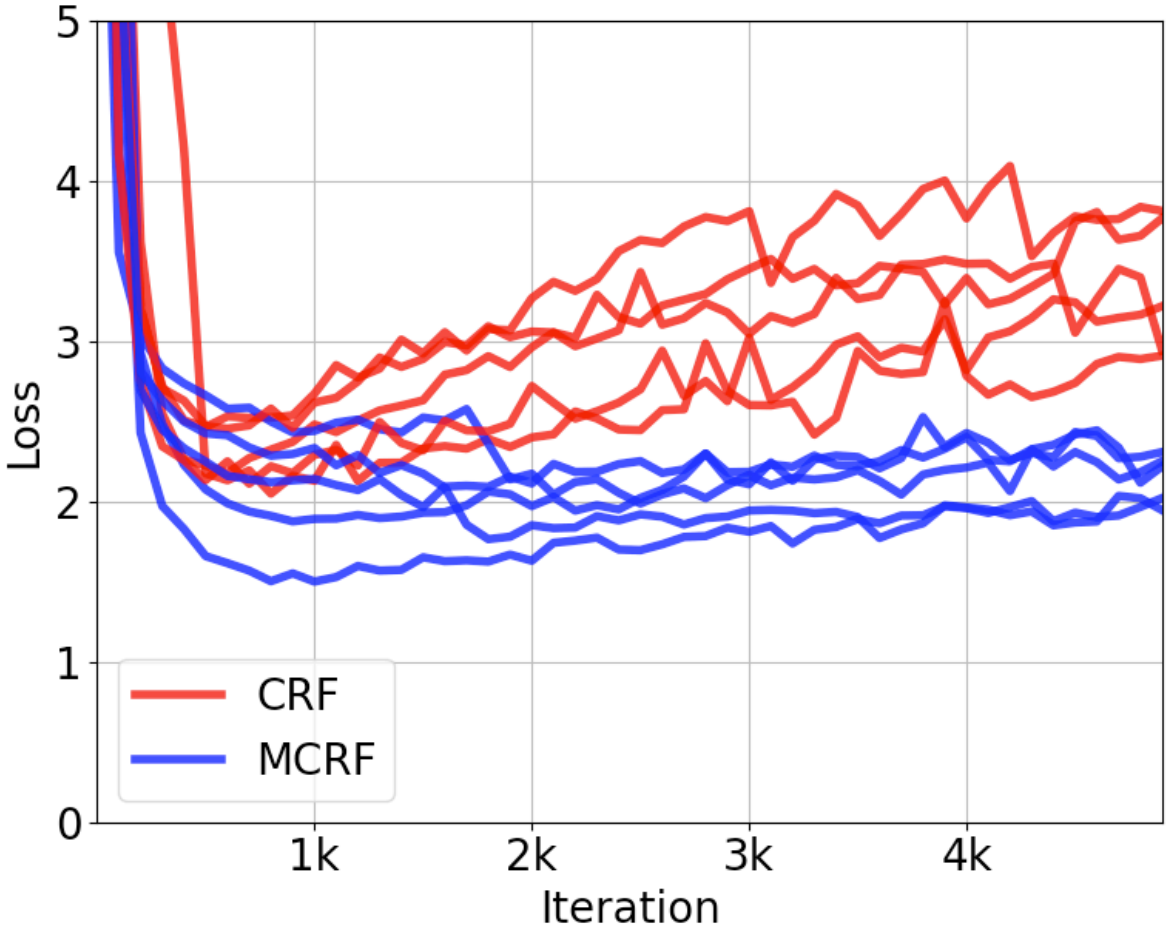


Figure: Curves of dev loss for CRF and MCRF.

Results on Slot Filling

Model	ATIS	SNIPS
(Goo et al., 2018)	95.4	89.3
(Li et al., 2018)	96.5	-
(Zhang et al., 2019)	95.2	91.8
(E et al., 2019)	95.8	92.2
(Siddhant et al., 2019)	95.6	93.9
BERT-tagger-retain	95.2 (92.9)	93.2 (92.1)
BERT-tagger-discard	95.6 (93.1)	93.5 (92.3)
BERT-CRF-retain	95.5 (93.5)	94.6 (93.7)
BERT-CRF-discard	95.8 (93.9)	95.1 (94.3)
BERT-MCRF-decoding	95.8 (93.9)	95.1 (94.4)
BERT-MCRF-training	**95.9** (94.4)	**95.3** (94.6)

Results on Chunking

Model	F1
ELMo (Peters et al., 2017)	96.4
CSE (Akbik et al., 2018)	96.7
GCDT (Liu et al., 2019)	97.3
BERT-tagger-retain	96.1 (95.7)
BERT-tagger-discard	96.3 (96.0)
BERT-CRF-retain	96.5 (96.2)
BERT-CRF-discard	96.6 (96.3)
BERT-MCRF-decoding	96.6 (96.4)
BERT-MCRF-training	**96.9** (96.5)

Ablation over Tagging Scheme (BIO v.s. BIOES)

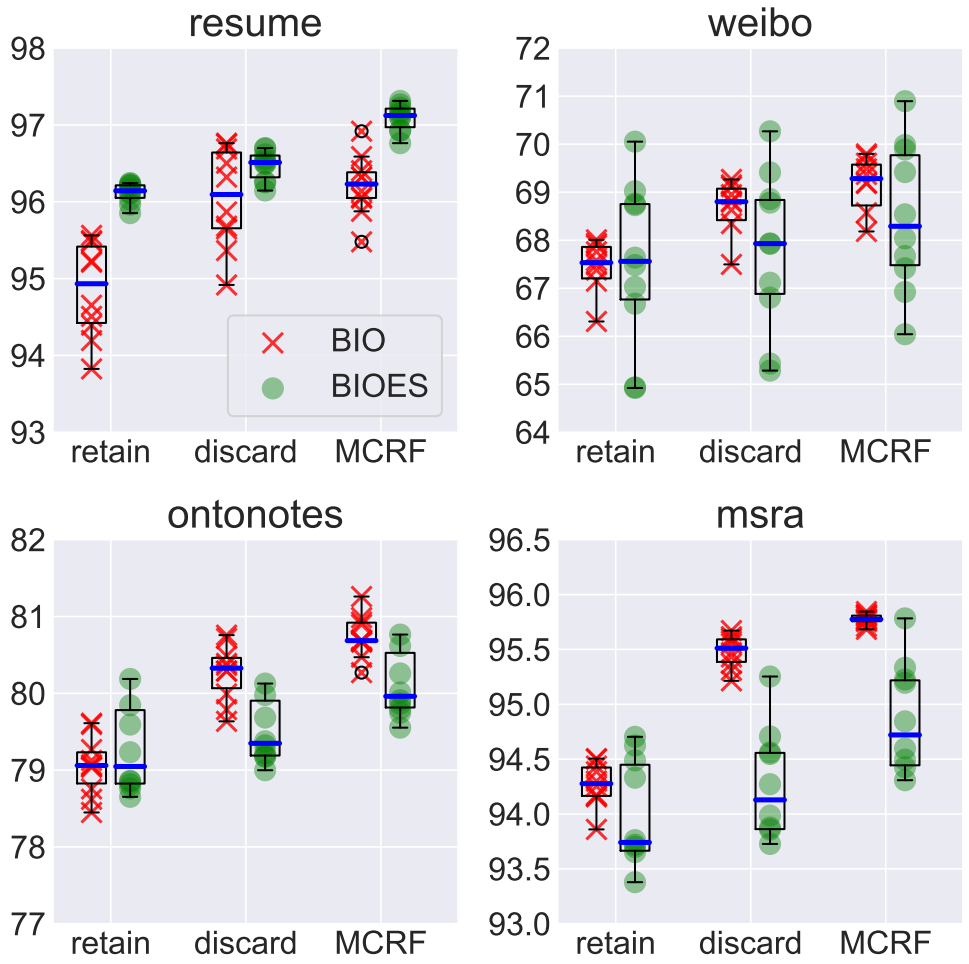


Figure: Ablation over the tagging scheme (BIO vs. BIOES). The F1-scores on the dev sets are plotted.

Ablation over Size of Training Corpus

	MSRA-full			MSRA-10%		
	ill.	F1	Δ	ill.	F1	Δ
retain	1.8%	94.2	1.6	2.4%	90.4	1.2
discard	-	95.4	0.5	-	90.7	0.9
MCRF	0%	95.8	-	0%	91.6	-
	Ontonotes-full			**Ontonotes-10%**		
	ill.	F1	Δ	ill.	F1	Δ
retain	4.2%	79.2	1.6	4.7%	78.7	1.2
discard	-	80.4	0.4	-	79.1	0.8
MCRF	0%	80.8	-	0%	79.9	-

Ablation over Model Architecture

Encoder	ill.	err.	CRF	MCRF	Δ
LSTM-1	3.1%	11.7%	82.2	83.2	1.0
LSTM-2	1.4%	8.3%	84.3	85.1	0.8
CNN + LSTM-1	0.4%	4.0%	94.1	94.3	0.2
CNN + LSTM-2	0.3%	2.3%	94.0	94.5	0.5
ELMo + LSTM-1	0.4%	3.3%	95.1	95.3	0.2
ELMo + LSTM-2	0.6%	5.5%	95.0	95.3	0.3
BERT	1.3%	12.5%	94.5	95.4	0.9
BERT + LSTM-1	1.0%	13.1%	94.7	95.3	0.6
BERT + LSTM-2	0.9%	10.3%	93.9	95.0	1.1

Summary of Contributions

1. To the best of our knowledge we are the first to show that in the neural-CRF framework the illegal path problem is intrinsic and may accounts for non-negligible proportion (up to 40%) of total errors.
2. We propose Masked Conditional Random Field (MCRF), an improved version of the CRF that is by design immune to the illegal paths problem. We also devise an algorithm for MCRF that requires only a few lines of code to implement.
3. We show in comprehensive experiments that MCRF performs significantly better than its CRF counterpart, and that its performance is on par with or better than more sophisticated models. We achieve new State-of-the-Arts in two Chinese NER datasets.

Thank You!