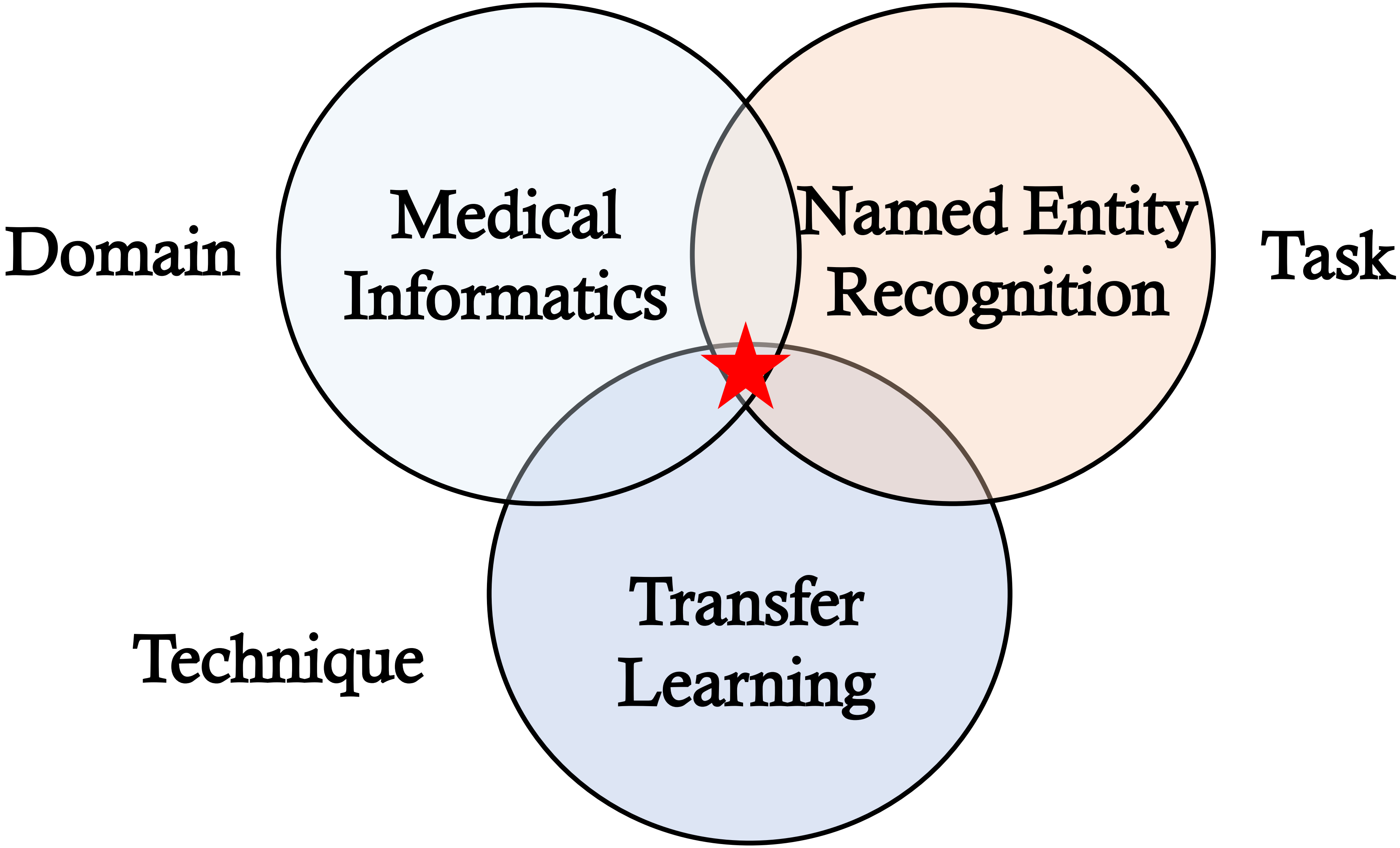
Label-Aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition

Zhenghui Wang[1], Yanru Qu[1], Liheng Chen[1], Jian Shen[1], Weinan Zhang[1], Shaodian Zhang[1,2], Yimei Gao[2], Gen Gu[2], Ken Chen[2], and Yong Yu[1]

[1] Shanghai Jiao Tong University

[2] 森亿智能 SYNYI·AI

What Do We Study?



Contents

- Background & Motivation
- Our Proposal
- Experiments & Results

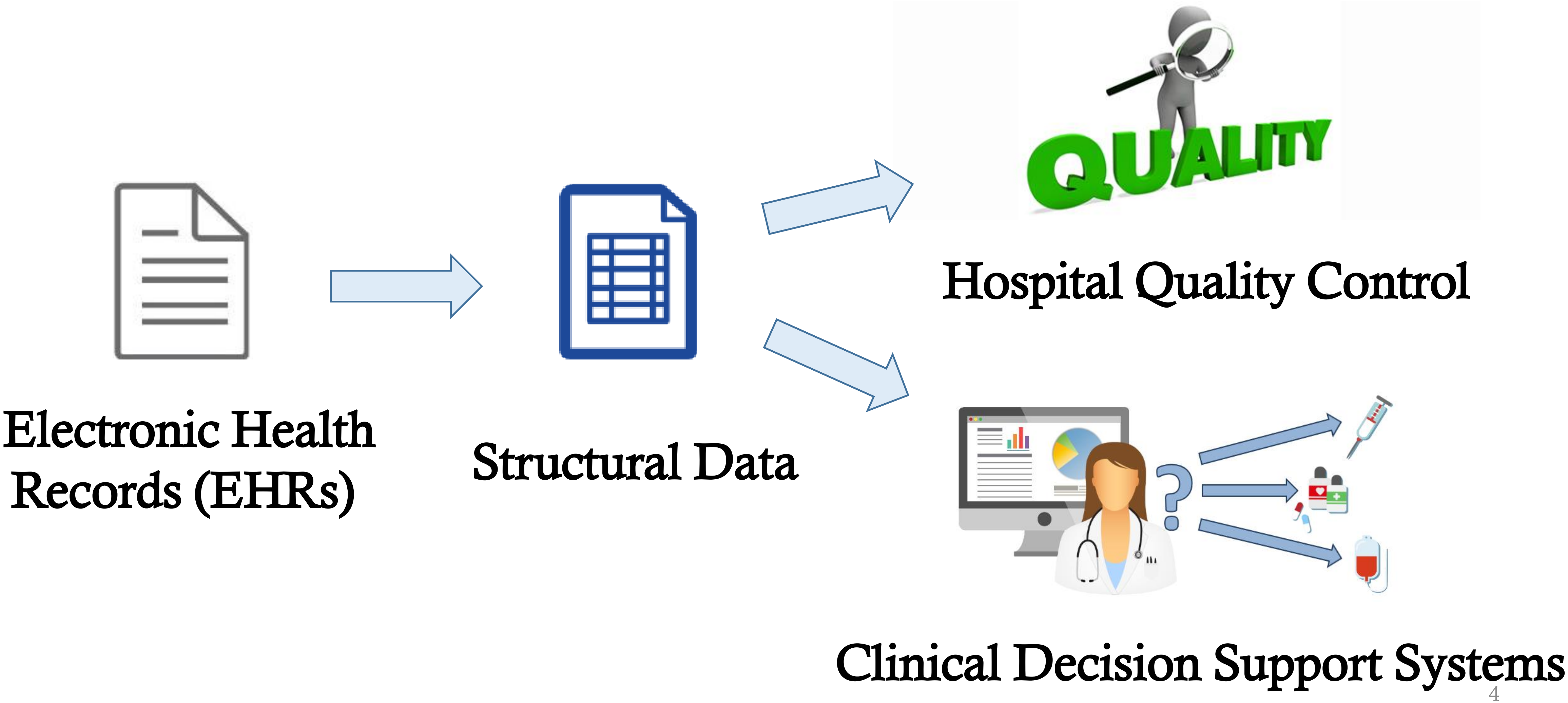
Medical Named Entity Recognition

Electronic Health Records (EHRs)

Hospital Quality Control

Clinical Decision Support Systems

Medical Named Entity Recognition



Medical Named Entity Recognition

But an MRI scan of the spine showed an L5 metastasis with a fracture

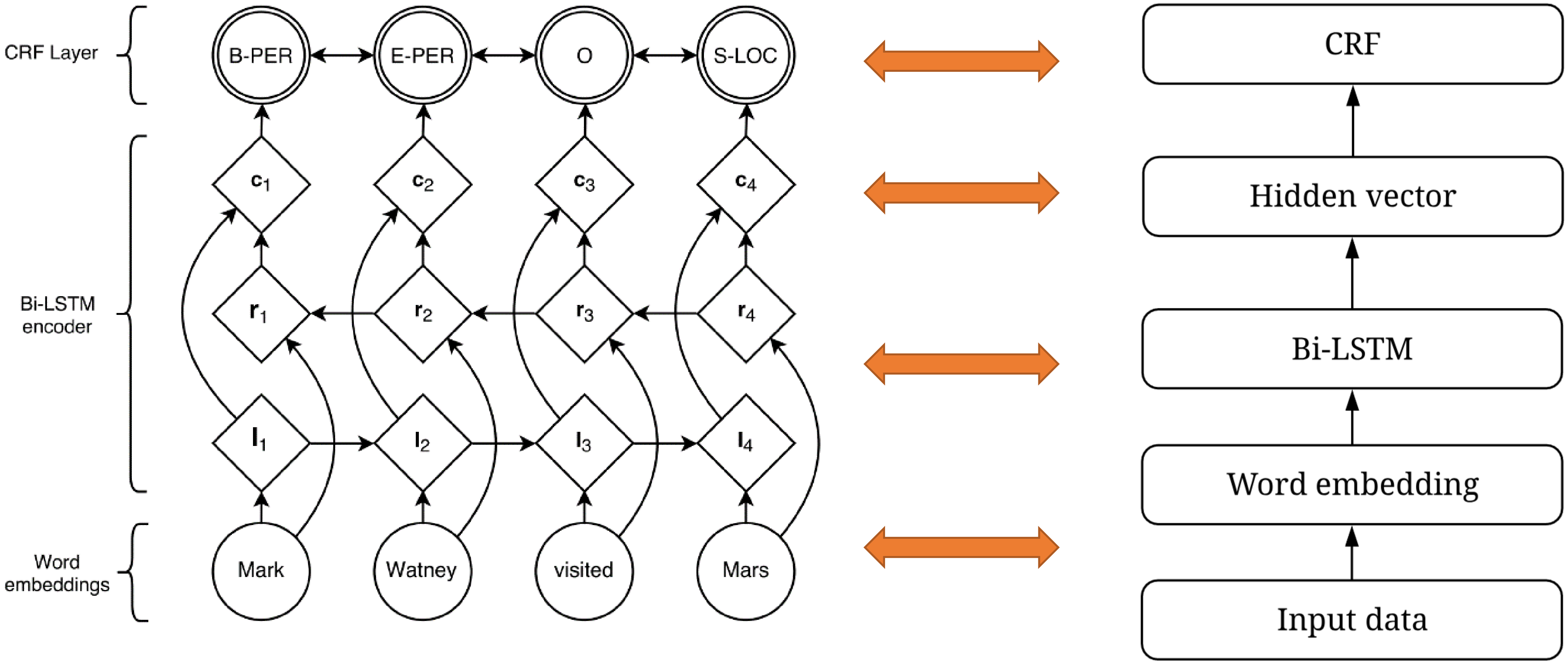
NER System

But	an	MRI	scan	of	the	spine	showed	an	L5	metastasis	with	a	fracture
O	B-T	I-T	I-T	I-T	I-T	I-T	O	B-P	I-P	I-P	O	B-P	I-P

$\underbrace{\text{an MRI scan of the spine}}_{\text{Test (T)}}$ $\underbrace{\text{an L5 metastasis}}_{\text{Problem (P)}}$ $\underbrace{\text{a fracture}}_{\text{Problem (P)}}$

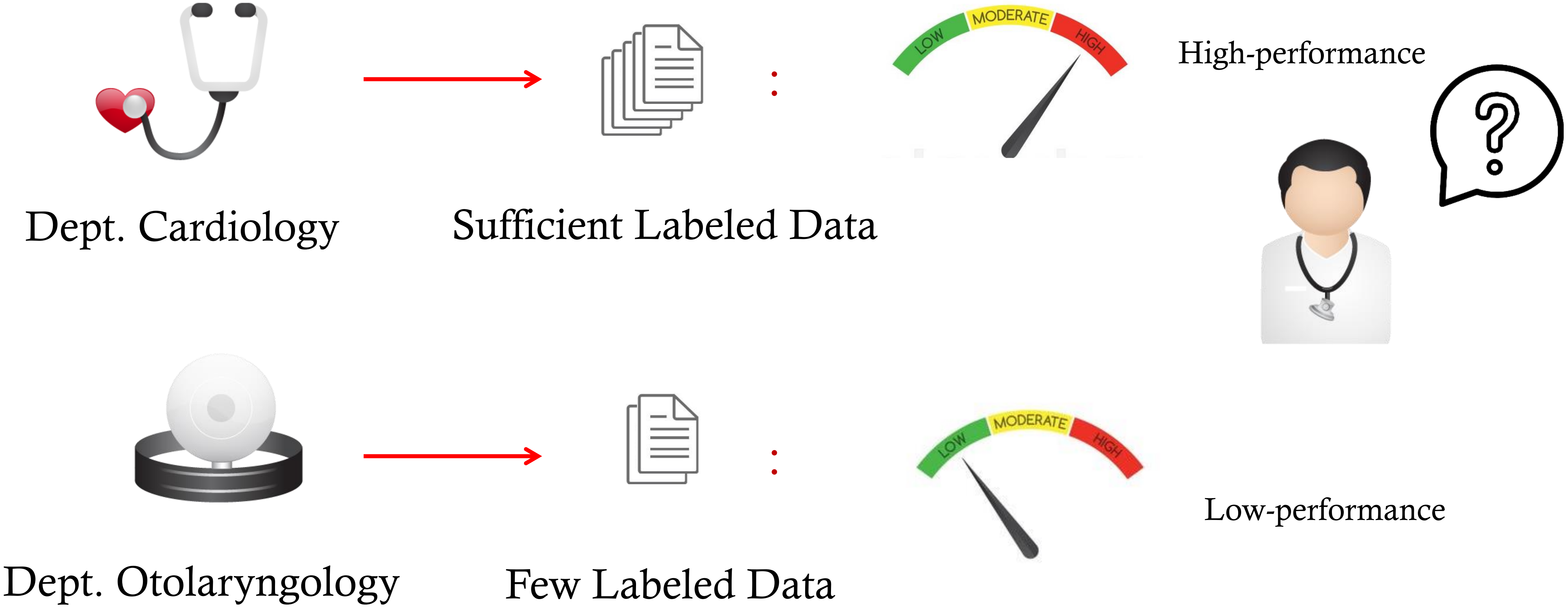
We focus on end-to-end NN-based models.

Bi-LSTM CRF



(Lample et al., 2016)

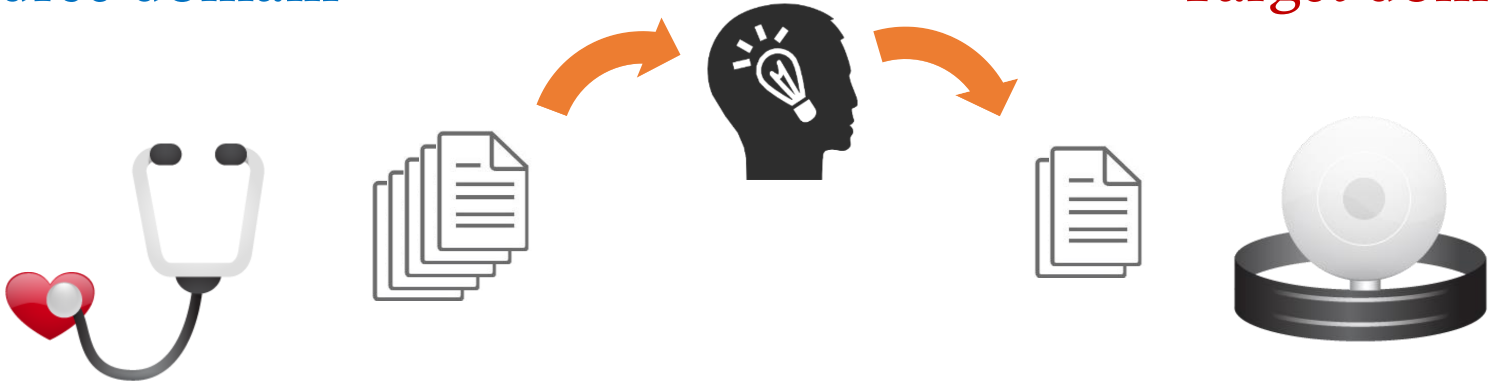
Challenge in Medical NER



Transfer Learning

Source domain

Target domain



Dept. Cardiology

Dept. Otolaryngology

Contents

- Background & Motivation
- Our Proposal
- Experiments & Results

Label-Aware Double Transfer Learning

- Feature representation transfer
- Parameter transfer

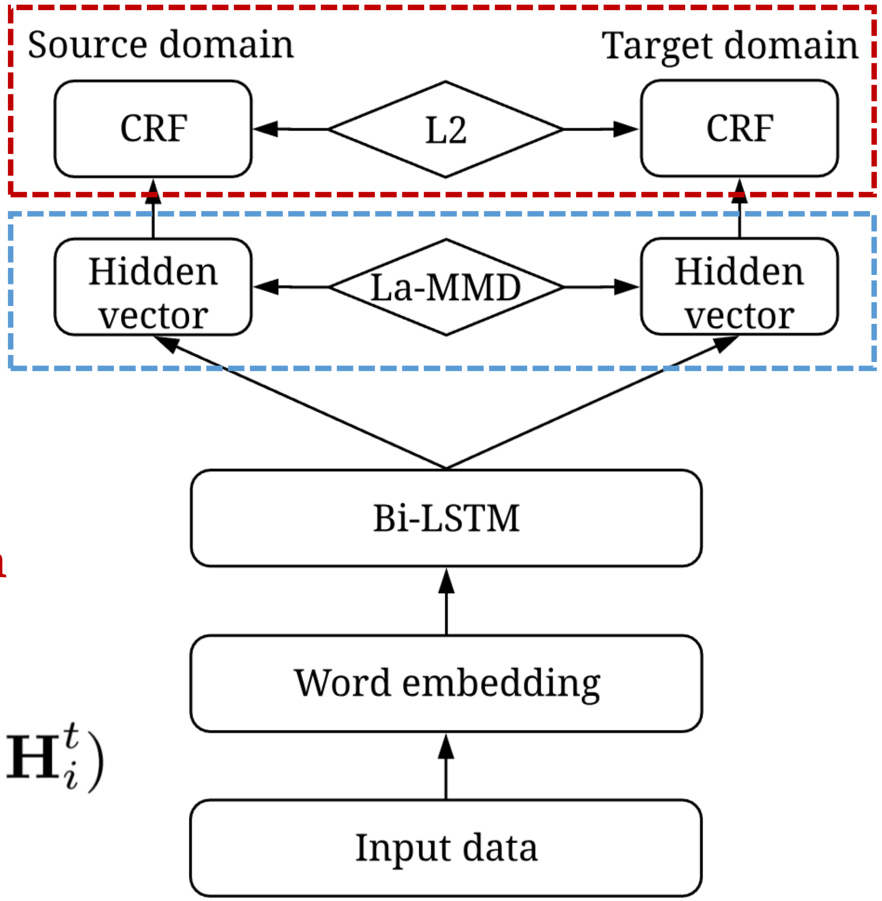
Parameter transfer loss

$$\mathcal{L} = \mathcal{L}_c + \boxed{\alpha\, \mathcal{L}_{\text{La-MMD}}} + \boxed{\beta\, \mathcal{L}_p} + \gamma\, \mathcal{L}_r$$

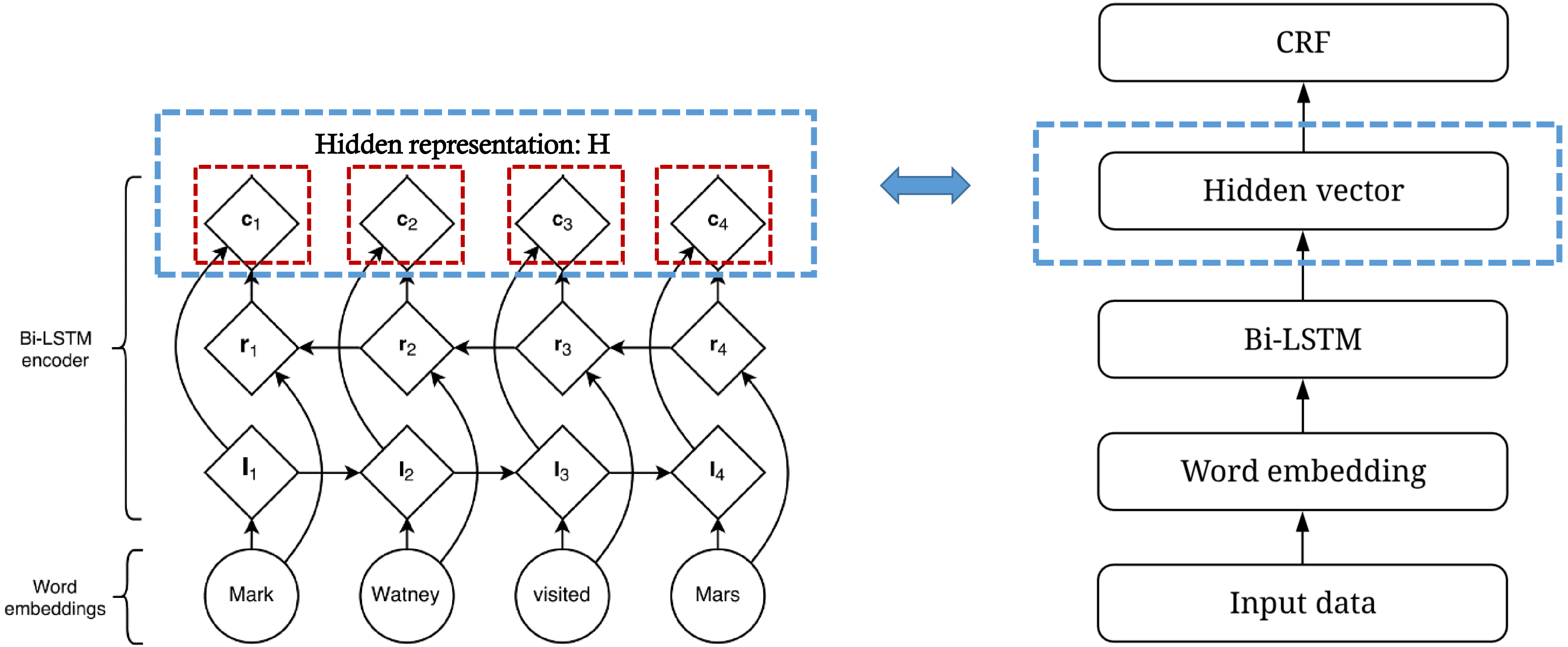
Feature representation transfer loss

Regularization

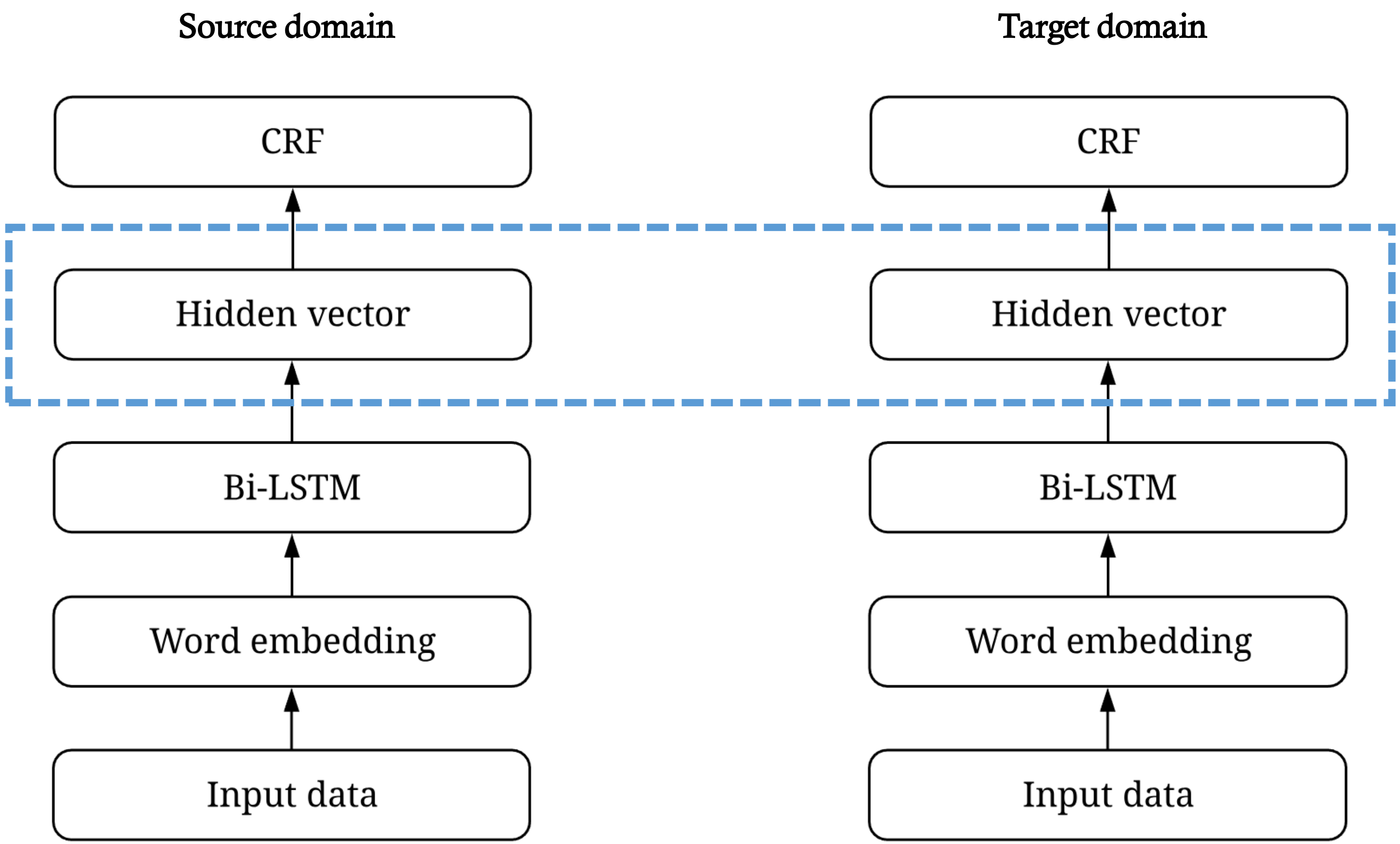
CRF loss: $$\mathcal{L}_c = -\frac{\varepsilon}{N^s}\sum_{i=1}^{N^s} \log p(\mathbf{y}_i^s|\mathbf{H}_i^s) - \frac{1-\varepsilon}{N^t}\sum_{i=1}^{N^t} \log p(\mathbf{y}_i^t|\mathbf{H}_i^t)$$



Feature representation transfer

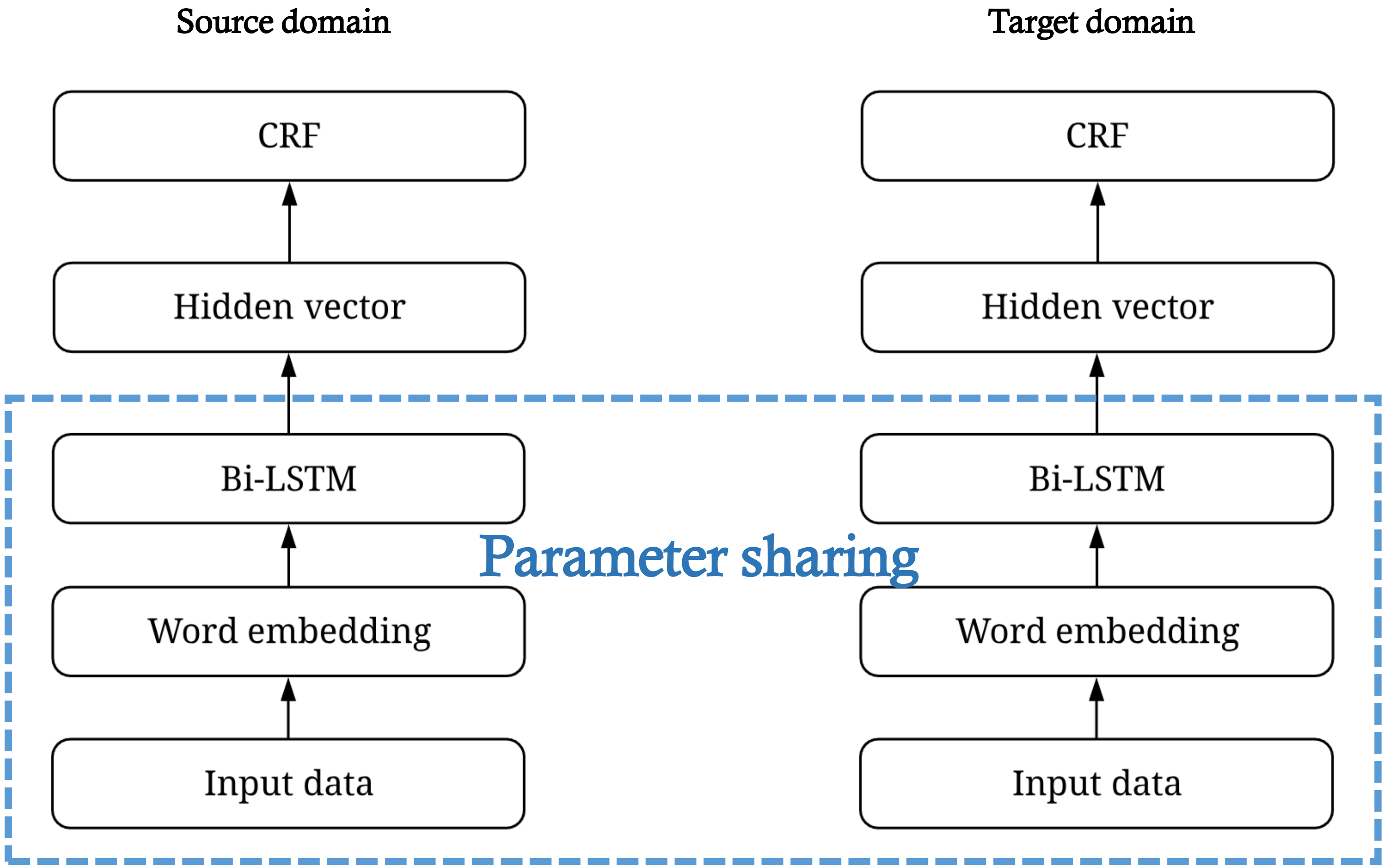


Feature representation transfer



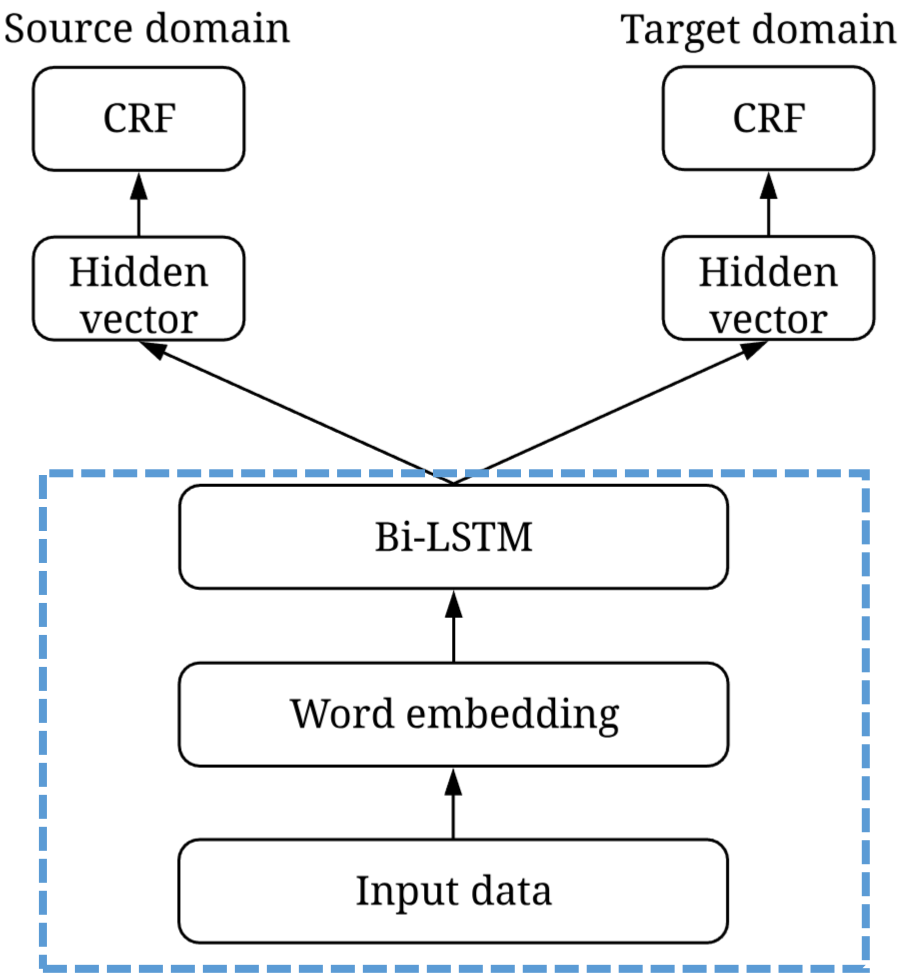
Feature representation transfer

- Parameter sharing



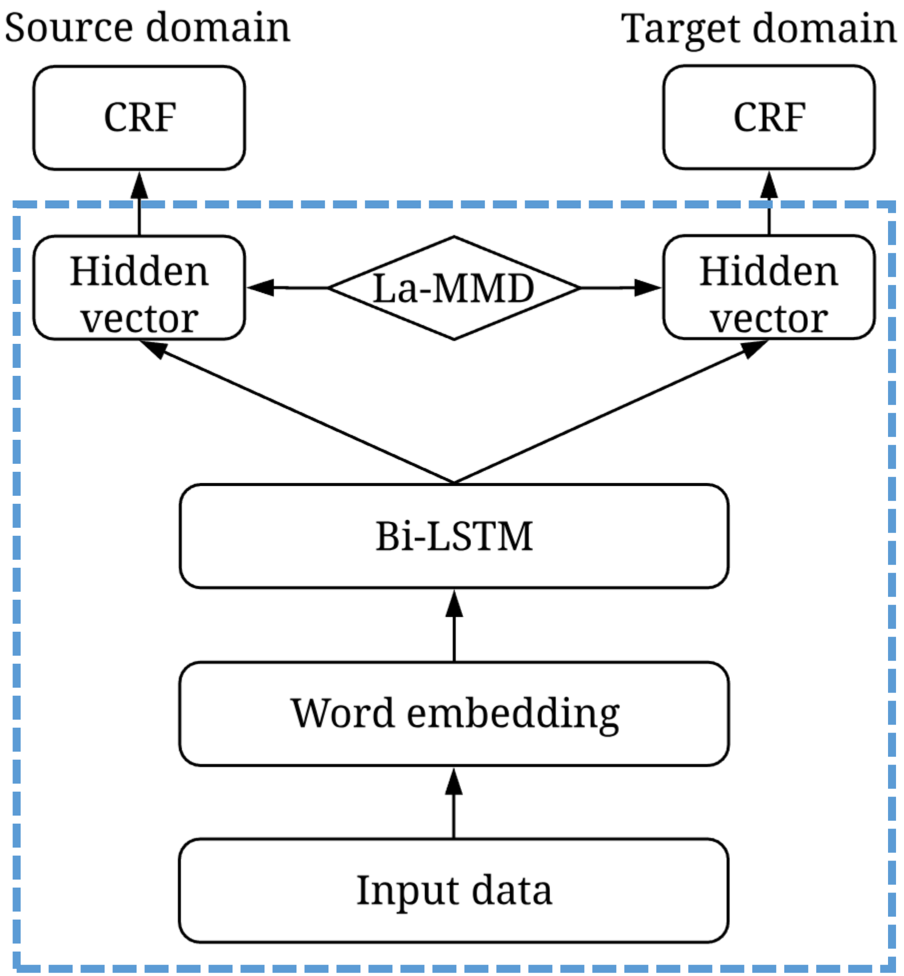
Feature representation transfer

- Parameter sharing



Feature representation transfer

- Label-aware MMD

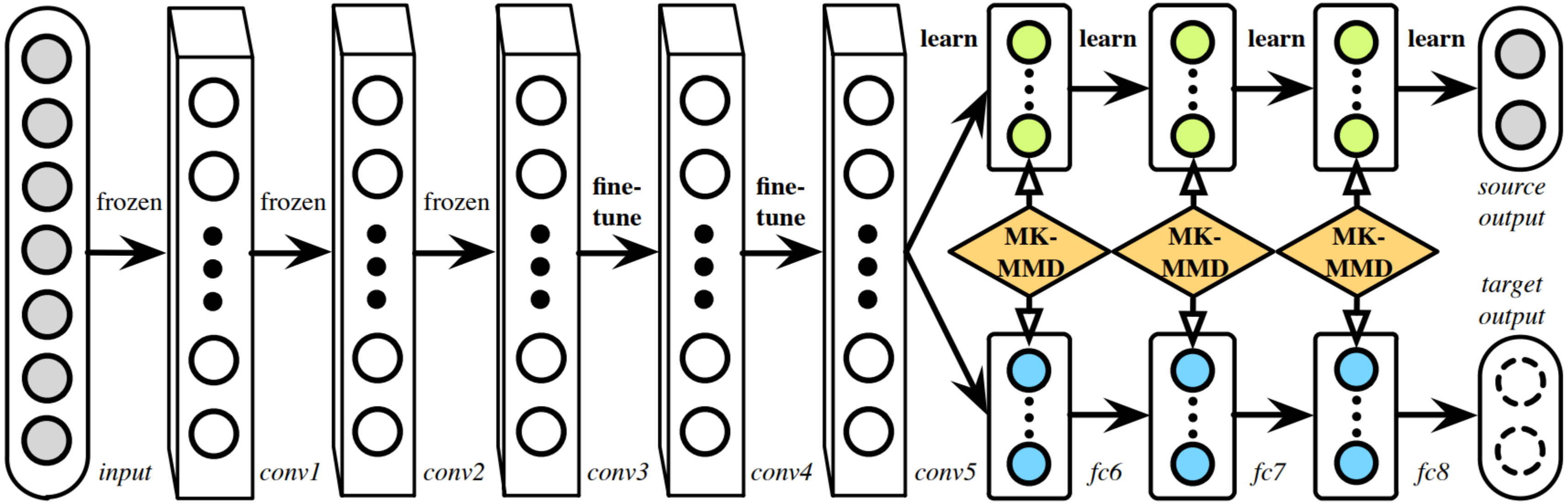


Maximum Mean Discrepancy (MMD)

- A non-parametric test statistic to measure the distribution discrepancy in terms of the distance between the kernel mean embeddings of two distributions p and q

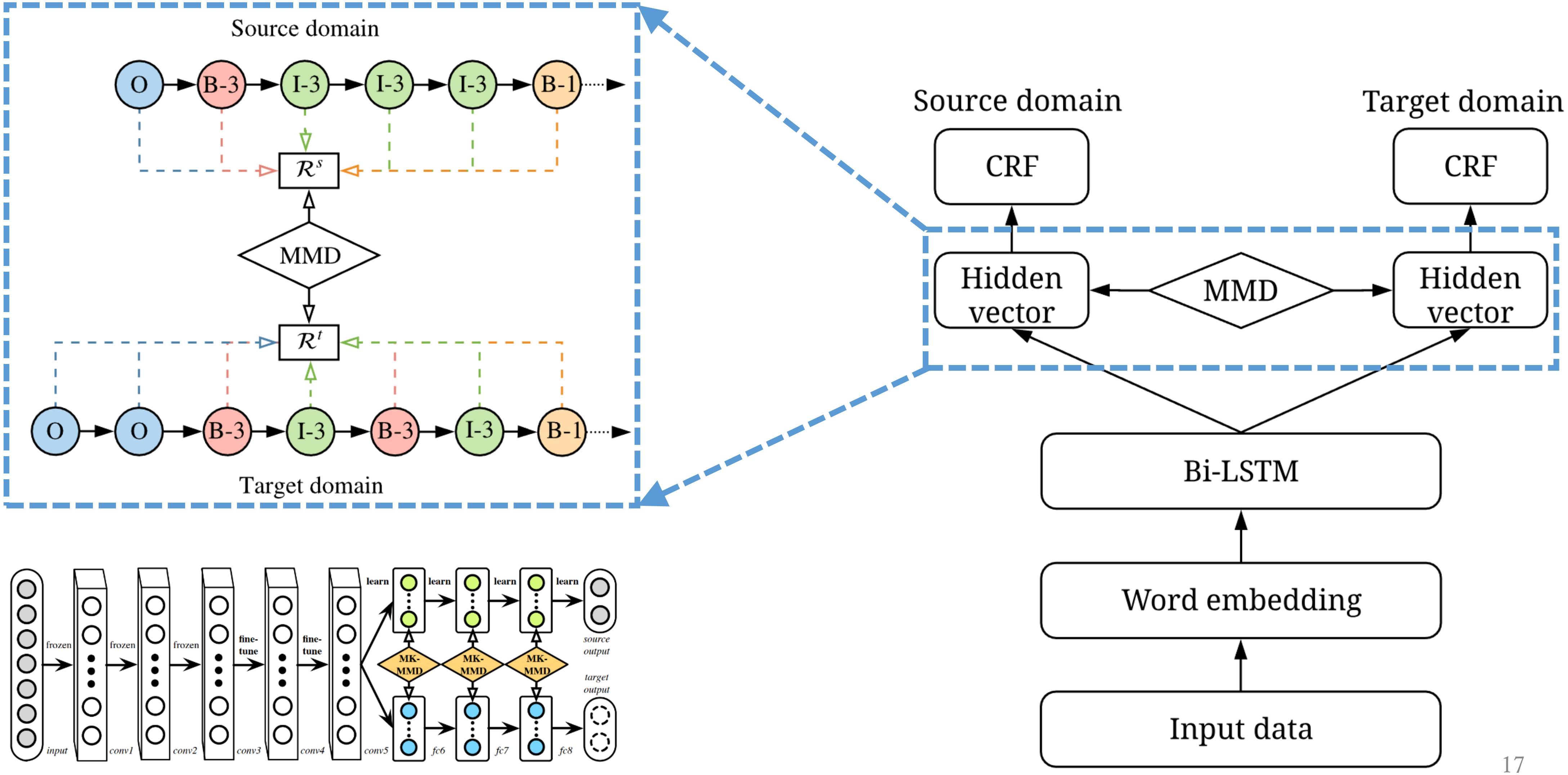
$$\mathrm{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])$$

- Deep Adaptation Network (DAN)



(Long et al., 2015)

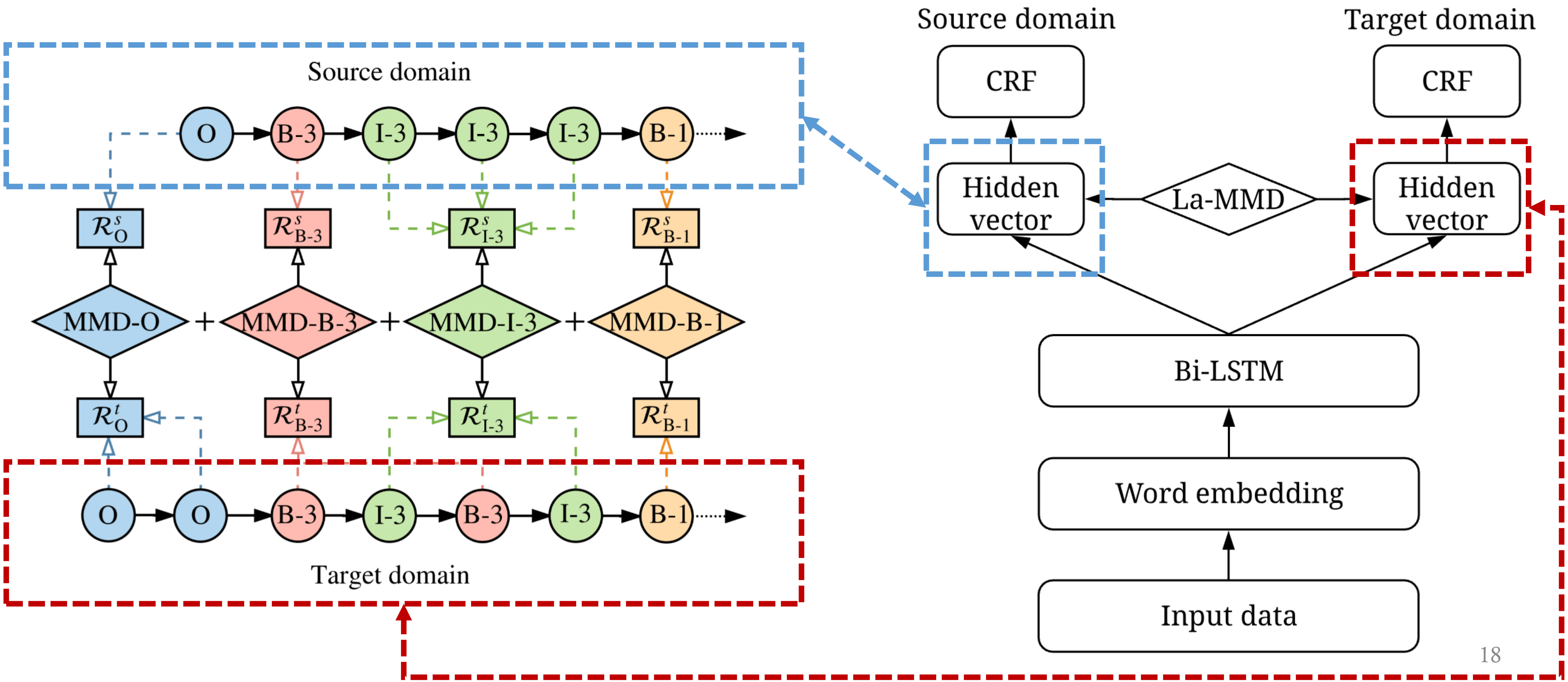
Vanilla MMD



Source domain
O
B-3
I-3
I-3
I-3
B-1
$\mathcal{R}^s$
MMD
$\mathcal{R}^t$
O
O
B-3
I-3
B-3
I-3
B-1
Target domain
input
frozen
conv1
frozen
conv2
frozen
conv3
fine-tune
conv4
fine-tune
conv5
learn
fc6
learn
fc7
learn
fc8
learn
MK-MMD
MK-MMD
MK-MMD
source output
target output
Source domain
Target domain
CRF
CRF
Hidden vector
MMD
Hidden vector
Bi-LSTM
Word embedding
Input data

Feature representation transfer

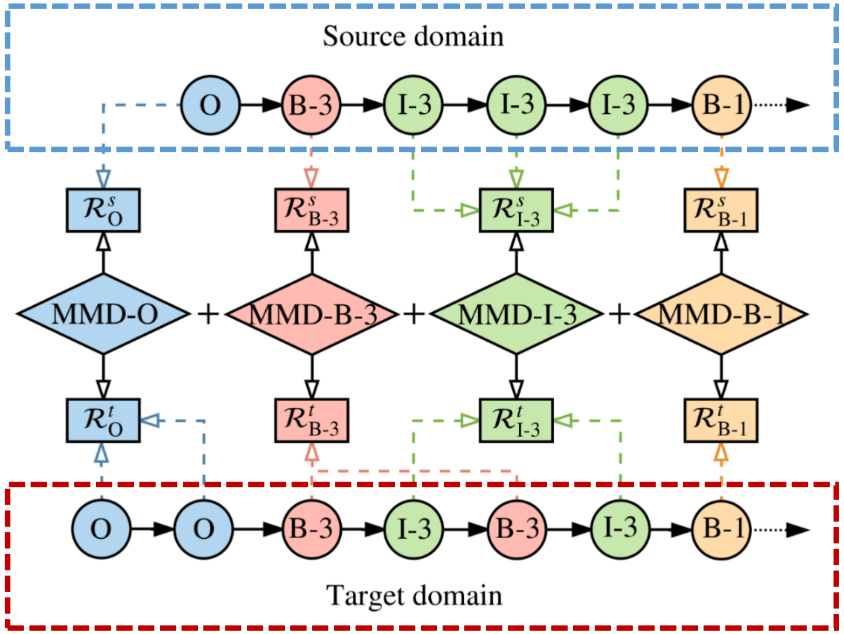
- Label-aware MMD



Feature representation transfer

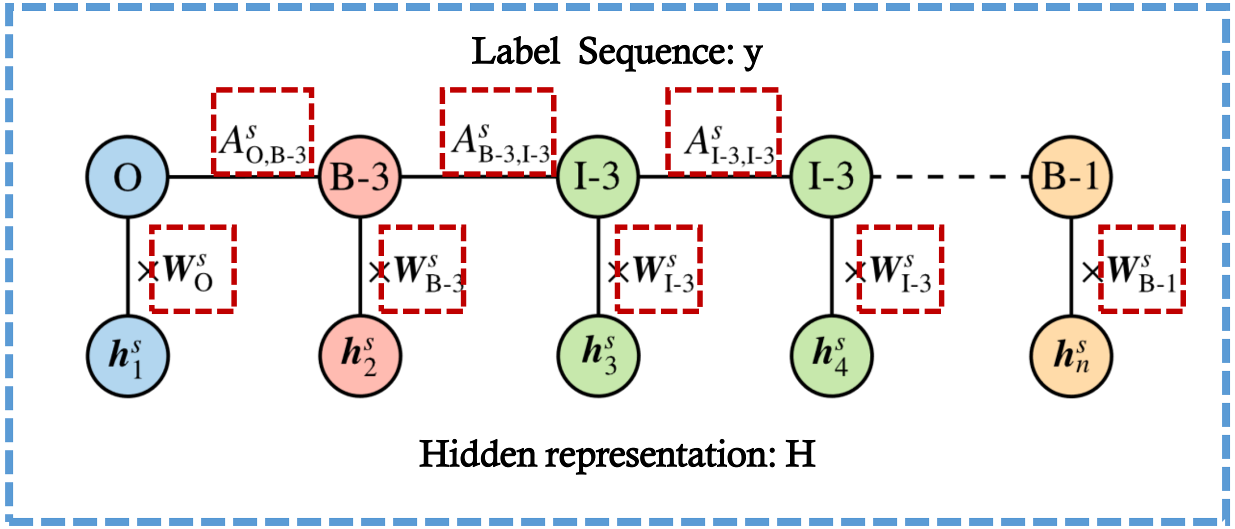
- Label-aware MMD

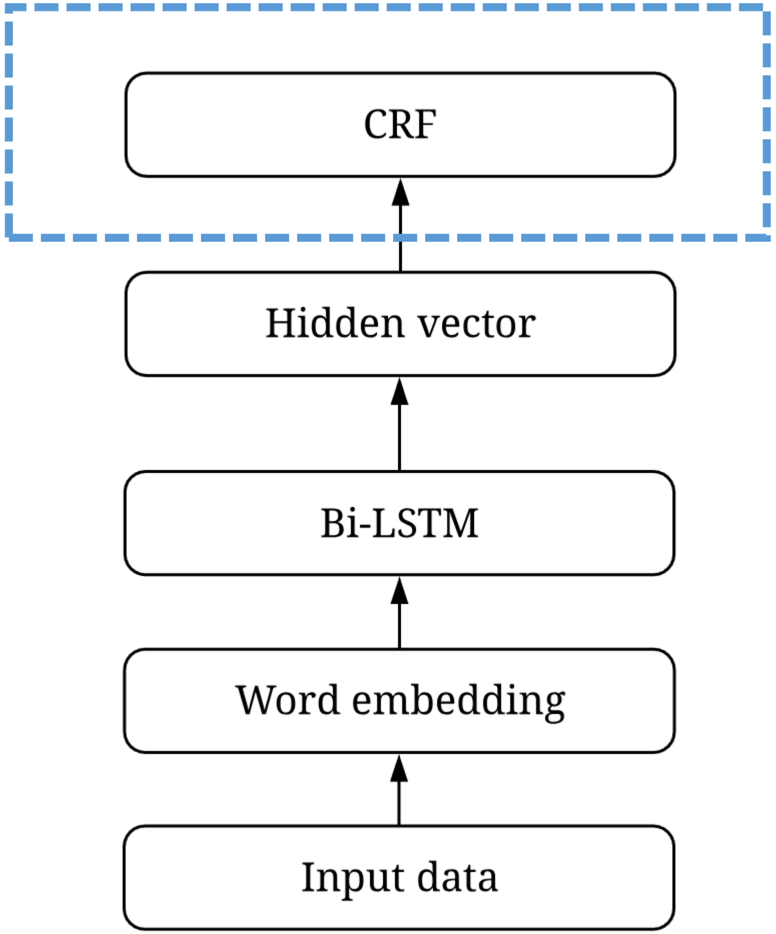
$$\mathcal{L}_{\text{La-MMD}} = \sum_{y \in \mathcal{Y}_v} \mu_y \cdot \mathrm{MMD}^2(\boxed{\mathcal{R}_y^s}, \boxed{\mathcal{R}_y^t}) \quad =$$



$$\mathrm{MMD}^2(\mathcal{R}_y^s, \mathcal{R}_y^t) = \frac{1}{(N_y^s)^2} \sum_{i,j=1}^{N_y^s} k(\boxed{\mathbf{h}_i^s, \mathbf{h}_j^s}) + \frac{1}{(N_y^t)^2} \sum_{i,j=1}^{N_y^t} k(\boxed{\mathbf{h}_i^t, \mathbf{h}_j^t}) \quad \frac{2}{N_y^s N_y^t} \sum_{i,j=1}^{N_y^s, N_y^t} k(\boxed{\mathbf{h}_i^s}, \boxed{\mathbf{h}_j^t})$$

Parameter transfer

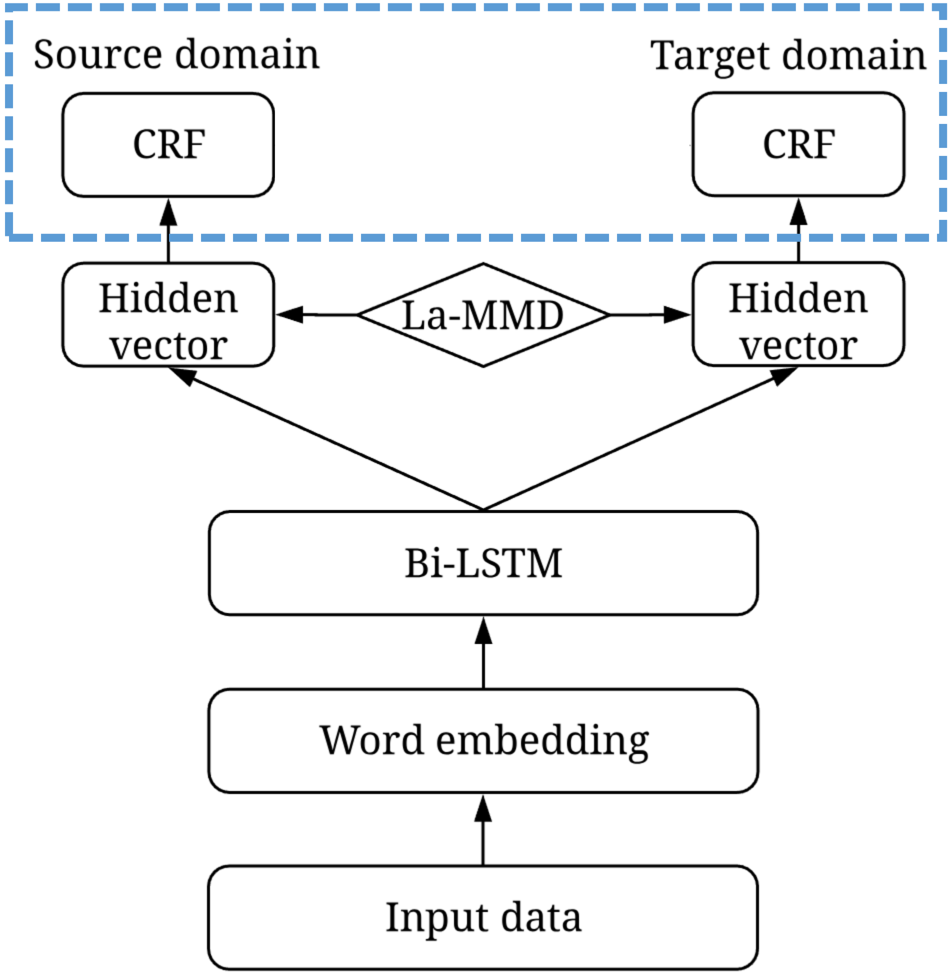




$$s_{\theta_c}(\mathbf{H}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{E}_{i,y_i} + \sum_{i=1}^{n-1} \mathbf{A}_{y_i, y_{i+1}}$$

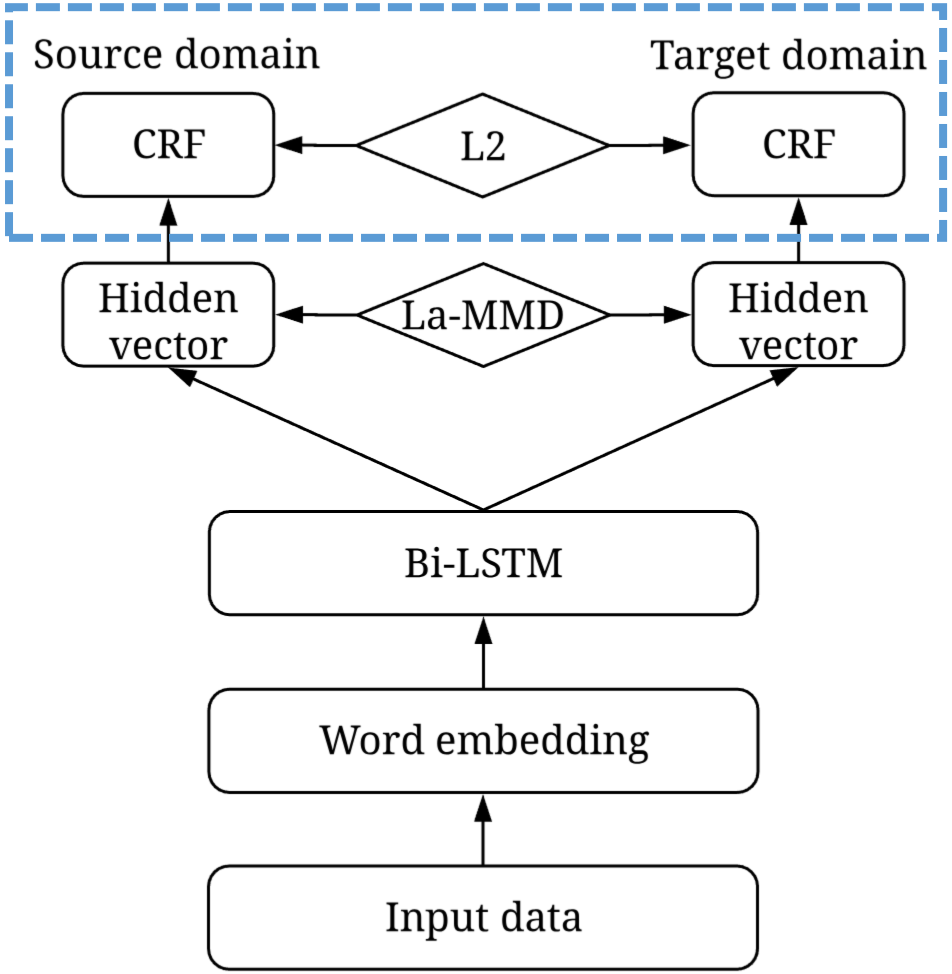
$$\begin{aligned} p_{\theta_c}(\mathbf{y}|\mathbf{H}) &= \exp\{s_{\theta_c}(\mathbf{H}, \mathbf{y})\}/Z(\mathbf{H}) \\ &= \exp\{s_{\theta_c}(\mathbf{H}, \mathbf{y})\} \Big/ \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{H})} \exp\{s_{\theta_c}(\mathbf{H}, \mathbf{y}')\} \end{aligned}$$

Parameter transfer



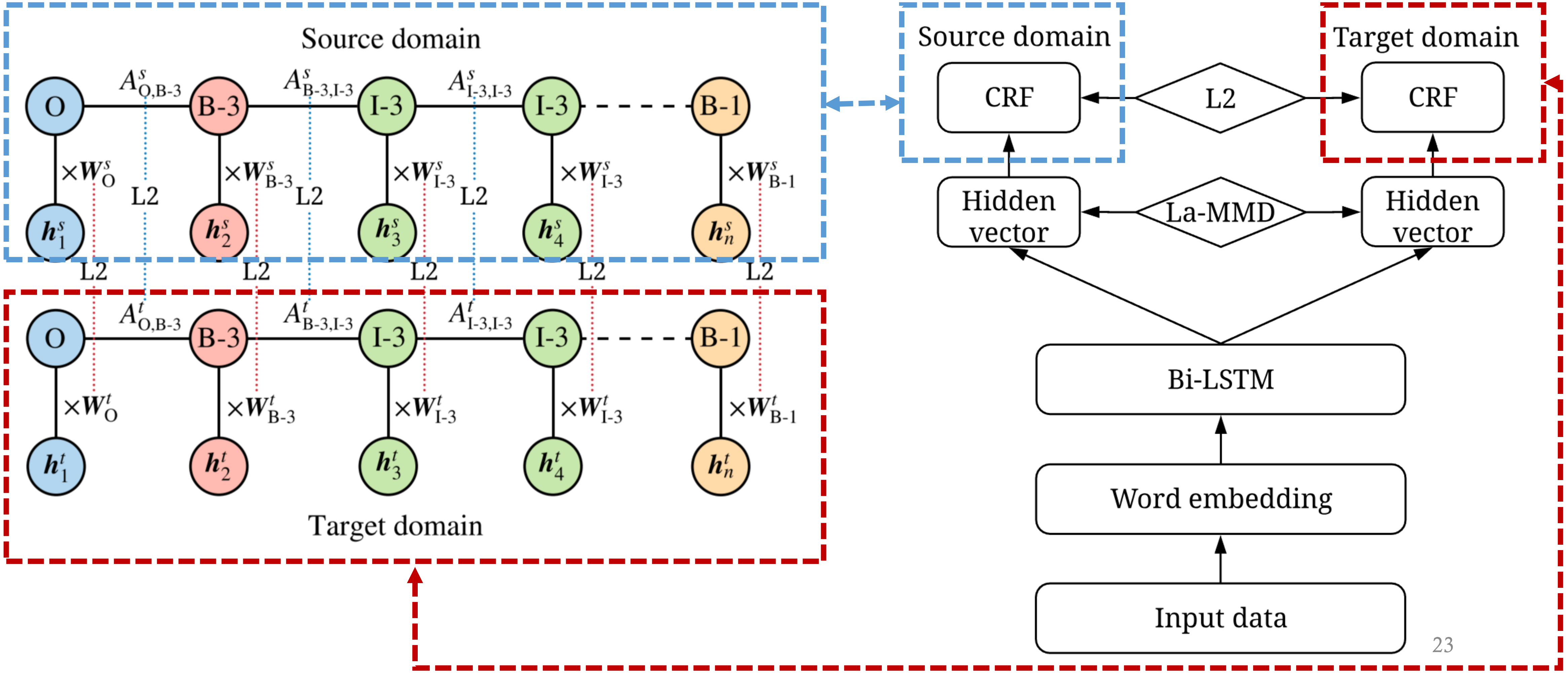
Parameter transfer

- L2 on CRF parameter



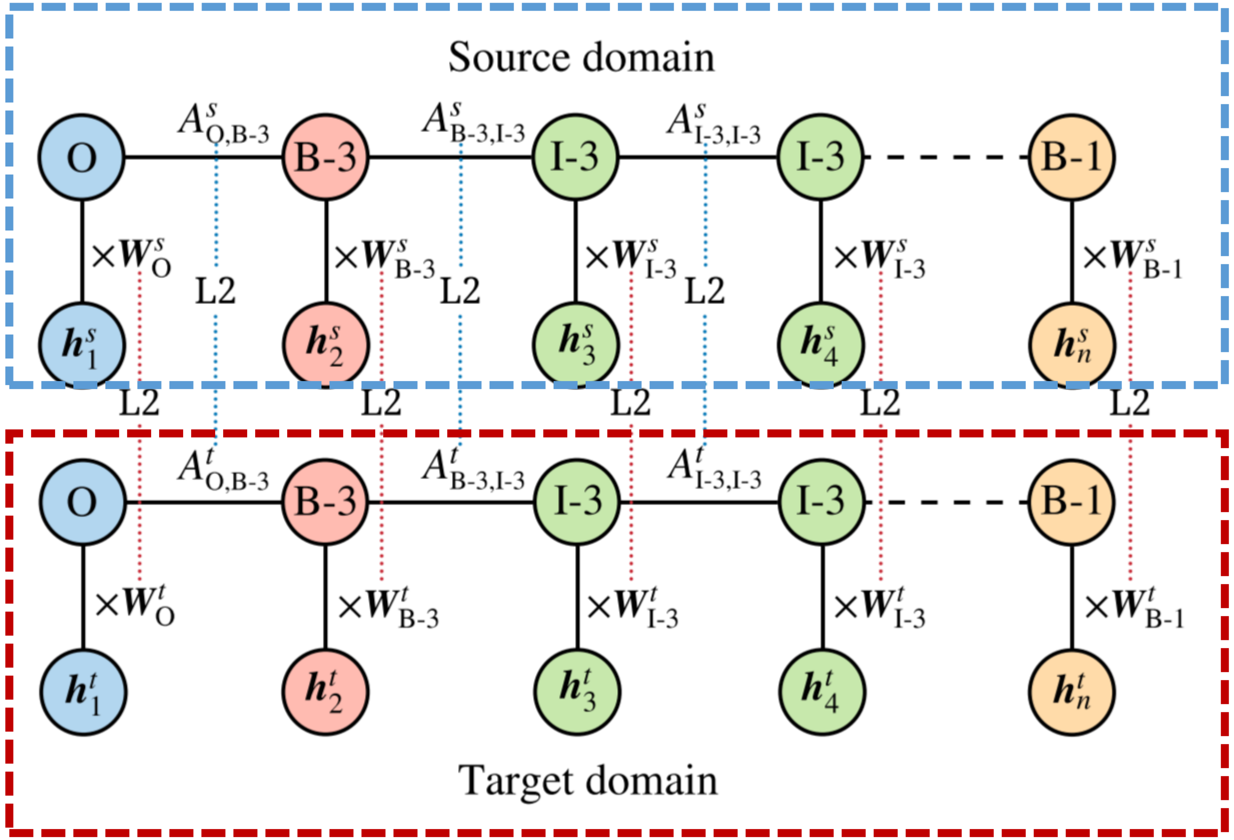
Parameter transfer

- L2 on CRF parameter



Parameter transfer

- L2 on CRF parameter

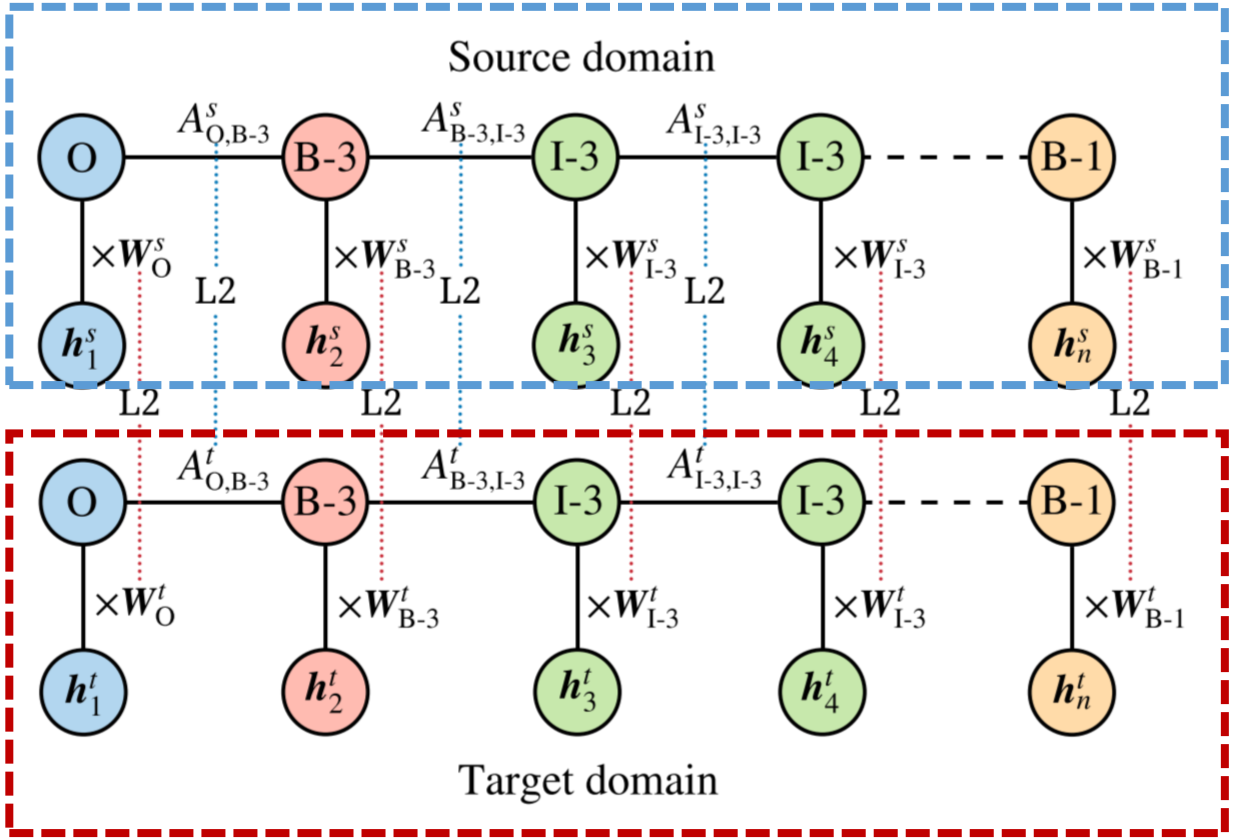


$$\mathcal{L}_p = \|\mathbf{W}^s - \mathbf{W}^t\|_2^2 + \|\mathbf{A}^s - \mathbf{A}^t\|_2^2$$

$$s_{\theta_c}(\mathbf{H}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{E}_{i,y_i} + \sum_{i=1}^{n-1} \mathbf{A}_{y_i,y_{i+1}}$$

Parameter transfer

- Bound



$$
\begin{aligned}
& D_{\mathrm{KL}}(p^s(\mathbf{y}|\mathbf{H})||p^t(\mathbf{y}|\mathbf{H})) \\
= & \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{H})} p^s(\mathbf{y}|\mathbf{H}) \log(\frac{p^s(\mathbf{y}|\mathbf{H})}{p^t(\mathbf{y}|\mathbf{H})}) \\
\leq & c(||\boxed{\mathbf{W}^s} - \boxed{\mathbf{W}^t}||_2^2 + ||\boxed{\mathbf{A}^s} - \boxed{\mathbf{A}^t}||_2^2)^{\frac{1}{2}}
\end{aligned}
$$

Label-Aware Double Transfer Learning

- Feature representation transfer
- Parameter transfer

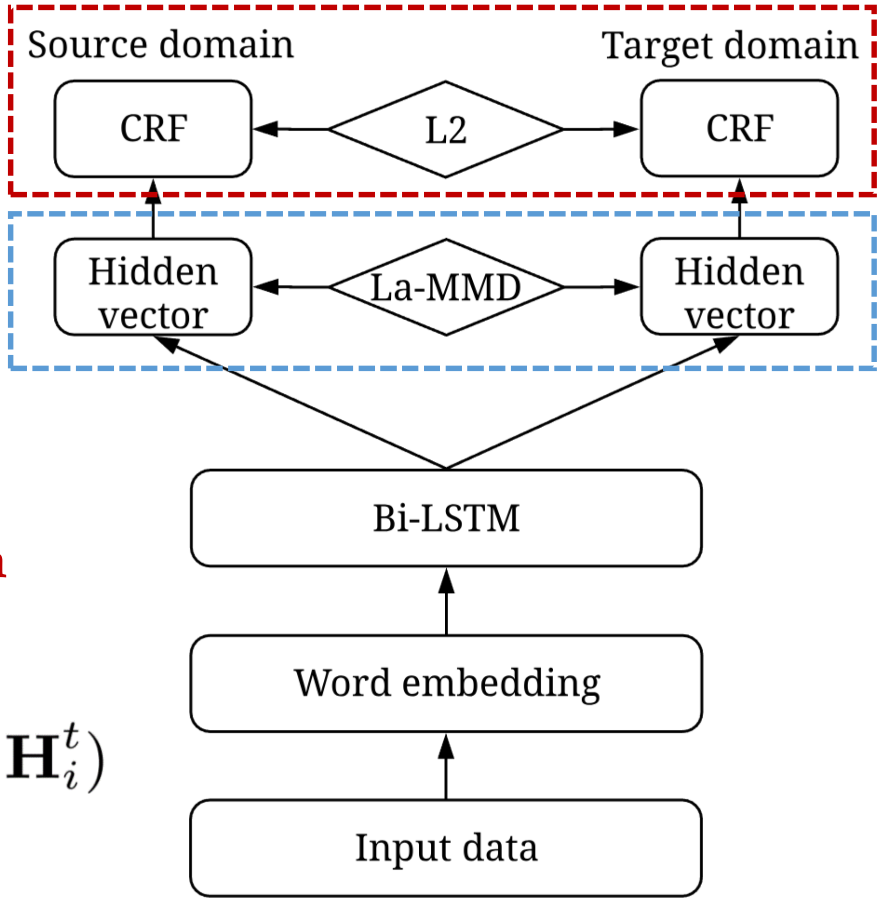
Parameter transfer loss

$$\mathcal{L} = \mathcal{L}_c + \boxed{\alpha\ \mathcal{L}_{\text{La-MMD}}} + \boxed{\beta\ \mathcal{L}_p} + \gamma\ \mathcal{L}_r$$

Feature representation transfer loss

Regularization

CRF loss: $\mathcal{L}_c = -\frac{\varepsilon}{N^s}\sum_{i=1}^{N^s}\log p(\mathbf{y}_i^s|\mathbf{H}_i^s) - \frac{1-\varepsilon}{N^t}\sum_{i=1}^{N^t}\log p(\mathbf{y}_i^t|\mathbf{H}_i^t)$



Contents

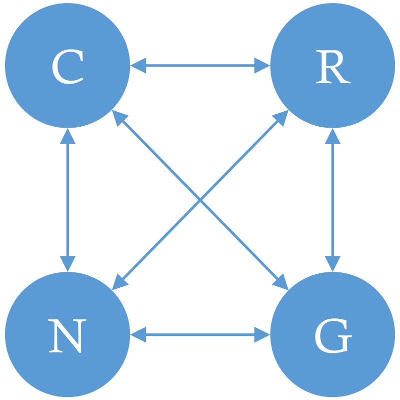
- Background & Motivation
- Our Proposal
- Experiments & Results

Experiments

- De-identified EHRs from 4 departments:

Department	# Train	# Dev	# Test
Cardiology (C)	3,004	601	601
Respiratory (R)	3,025	605	606
Neurology (N)	932	187	187
Gastroenterology (G)	1,517	303	304

- 12 transfer tasks:



Experiments

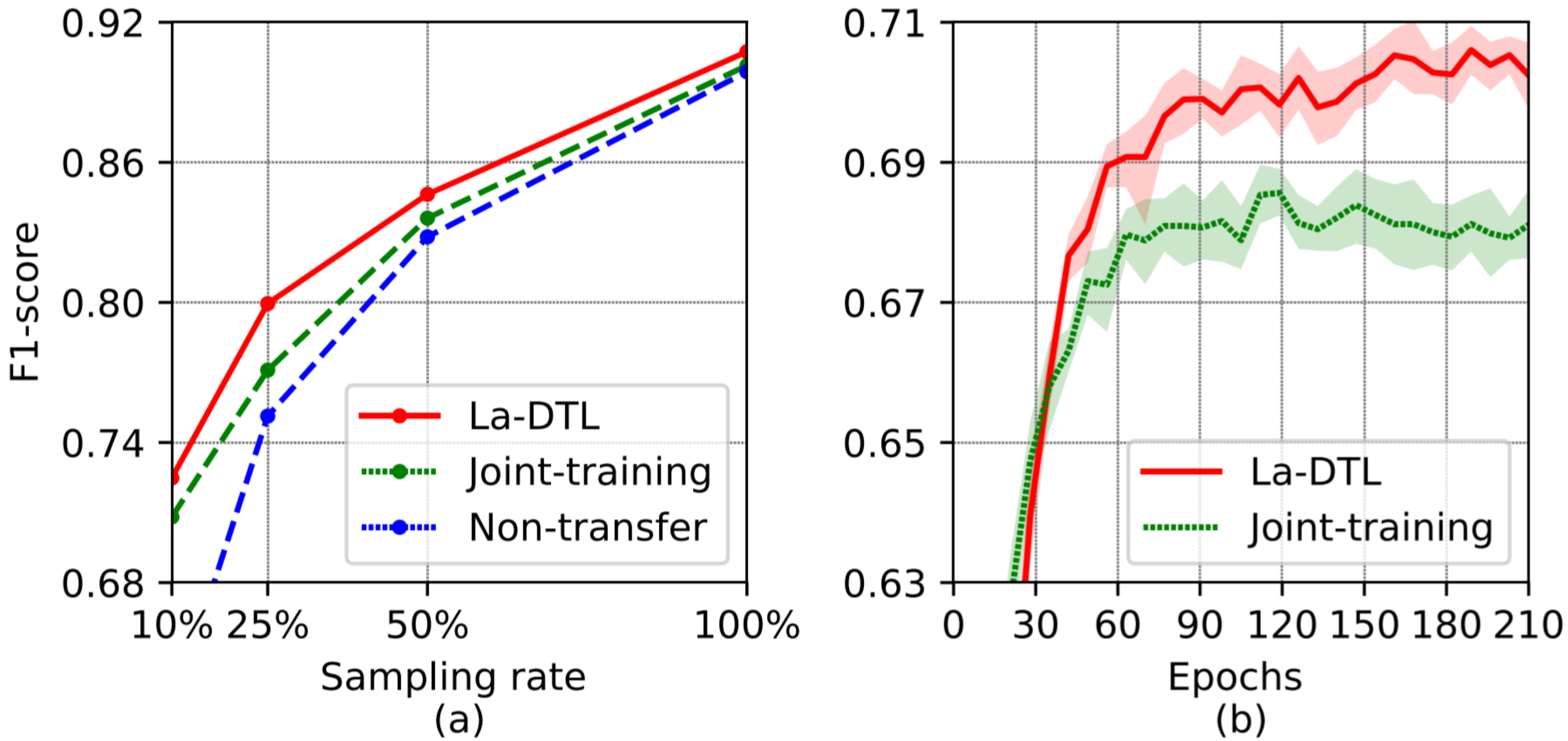
- 12 transfer tasks
  - 2.62% to 6.70% average F1-score improvement

Method	C→R	C→N	C→G	R→C	R→N	R→G	N→C	N→R	N→G	G→C	G→R	G→N	AVG
Non-transfer	67.20	54.51	49.01	65.63	54.51	49.01	65.63	67.20	49.01	65.63	67.20	54.51	59.09
Linear projection (Peng and Dredze, 2017)	69.01	67.02	57.40	69.79	65.87	57.71	67.70	68.77	51.33	68.00	69.65	61.12	64.45
Domain mask (Peng and Dredze, 2017)	70.76	63.97	58.62	70.18	64.27	58.16	67.93	69.89	56.18	68.87	69.89	63.49	65.18
CD-learning (He and Sun, 2017)	71.38	64.01	56.72	72.17	64.91	58.14	68.99	71.13	56.27	70.17	71.76	62.06	65.64
Re-training (Lee et al., 2017)	72.45	70.55	59.58	72.56	68.59	60.94	69.60	70.08	56.58	70.14	71.90	66.01	67.42
Joint-training (Yang et al., 2017)	69.82	70.49	63.52	71.45	67.03	67.71	70.96	71.43	60.54	69.68	71.55	68.15	68.53
La-MMD	73.08	69.48	59.86	72.53	70.28	60.16	71.31	73.04	57.94	69.80	73.99	67.19	68.22
CRF-L2	73.34	71.52	60.17	72.43	69.72	67.61	69.76	71.54	59.96	69.75	71.82	67.30	68.74
MMD-CRF-L2	73.05	72.35	60.80	72.65	69.87	66.82	70.25	71.75	58.98	70.48	73.98	67.43	69.03
La-DTL	**73.59**†	**72.91**†	**64.60**†	**73.88**†	**73.01**†	**70.17**†	**73.08**†	**73.11**†	**62.14**†	**71.61**†	**74.21**†	**71.49**†	**71.15**

(Peng and Dredze, 2017; He and Sun, 2017; Lee et al., 2017; Yang et al., 2017)

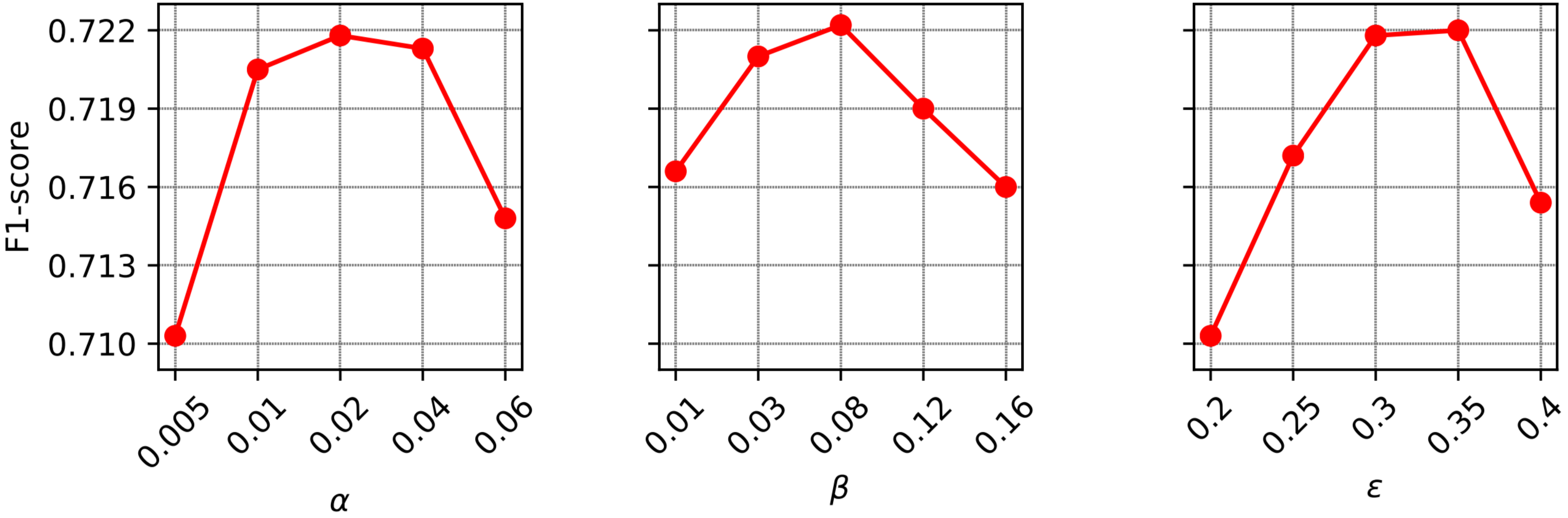
Experiments

- (a) Different target domain Sampling rate on C→R
- (b) Results of 10 trials on C→R (Cardiology → Respiratory)



Experiments

- Hyperparameter Study on C→R (Cardiology → Respiratory)



$\alpha \, \mathcal{L}_{\text{La-MMD}}$

Feature representation transfer

$\beta \, \mathcal{L}_p$

Parameter transfer

$$\mathcal{L}_c = -\frac{\varepsilon}{N^s} \sum_{i=1}^{N^s} \log p(\mathbf{y}_i^s | \mathbf{H}_i^s) - \frac{1-\varepsilon}{N^t} \sum_{i=1}^{N^t} \log p(\mathbf{y}_i^t | \mathbf{H}_i^t)$$

Balance CRF loss between source/target

Experiments on Social Media Domain

- SighanNER → WeiboNER

Method	F1-score
Non-transfer	54.78
Linear projection (Peng and Dredze, 2017)*	56.40
Linear projection (Peng and Dredze, 2017)	56.99
Domain mask (Peng and Dredze, 2017)*	56.80
Domain mask (Peng and Dredze, 2017)	56.32
CD-learning (He and Sun, 2017)*	52.05
CD-learning (He and Sun, 2017)	56.46
Re-training (Lee et al., 2017)	55.36
Joint-training (Yang et al., 2017)	56.80
La-DTL	**57.74**

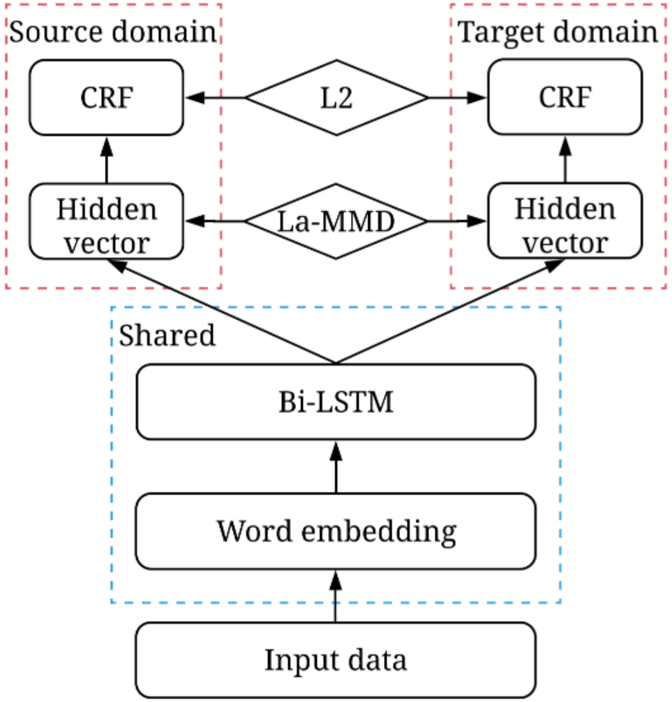
- CoNLL 2003 → TwitterNER

Method	F1-score
Non-transfer	34.65
Joint-training (Yang et al., 2017)*	43.24
La-DTL	**45.71**

Reference

- Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.
- Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[J]. arXiv preprint arXiv:1502.02791, 2015.
- Peng N, Dredze M. Multi-task Domain Adaptation for Sequence Tagging[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP. 2017: 91-100.
- He H, Sun X. A Unified Model for Cross-Domain and Semi-Supervised Named Entity Recognition in Chinese Social Media[C]//AAAI. 2017: 3216-3222.
- Lee J Y, Dernoncourt F, Szolovits P. Transfer Learning for Named-Entity Recognition with Neural Networks[J]. arXiv preprint arXiv:1705.06273, 2017.
- Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent networks[J]. arXiv preprint arXiv:1703.06345, 2017.

Label-Aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition



Thank You

- Feature representation transfer
- Parameter transfer

Zhenghui Wang
felixwzh AT apex.sjtu.edu.cn
zhenghuiwang.net