Multi-Specialty Domain Adaptation for Chinese Medical Named Entity Recognition

Zhucong Li, Baoli Zhang, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu

National Laboratory of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences

Beijing Unisound Information Technology Co., Ltd

医疗命名实体识别面临的重要挑战：医疗领域标注数据缺乏

医疗数据难以获取

医疗数据标注难度大

医疗领域标注数据缺乏

领域自适应

联邦学习

主动学习

……

为什么需要领域自适应？

当为了在某个领域(Target Domain)训练一个数据驱动的机器学习模型时，我们如果缺少这个Target Domain的标注数据，可以利用其他领域(Source Domain)已有数据进行训练。但是由于现在的模型归纳能力有限，通常在out-of-domain的数据上的性能会出现非常明显的下降。

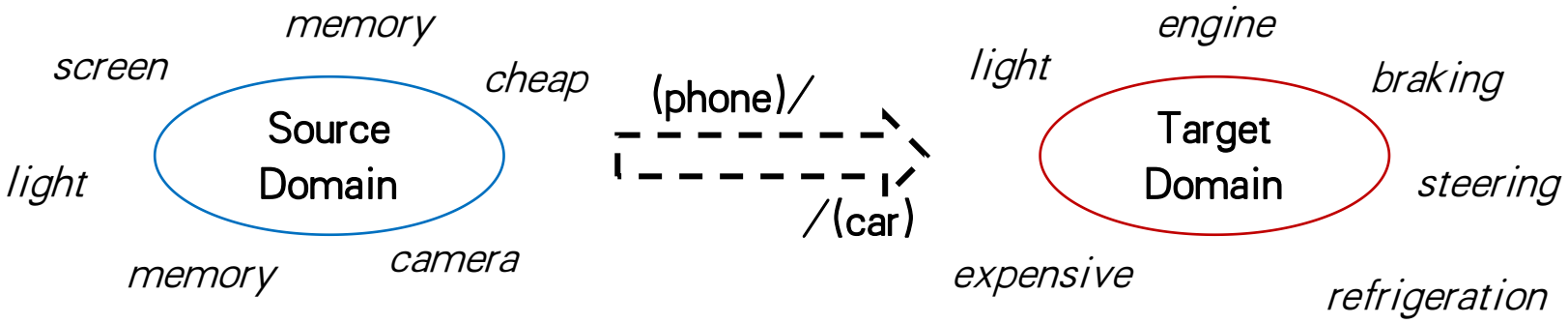
背景

例子

由于不同领域数据的分布不同，特性不同等等，模型在跨领域测试时往往会因为领域差异，使得数据驱动模型的能力被削弱。比如评论情感分析任务中(积极/消极)，同一个词在不同领域可能有不同的情感极性。

a（soruce domain==phone)：这部手机的重量很*轻* （``*轻*''是积极）

b（target domain==car)：这辆车的重量很*轻* （``*轻*''是消极）



背景

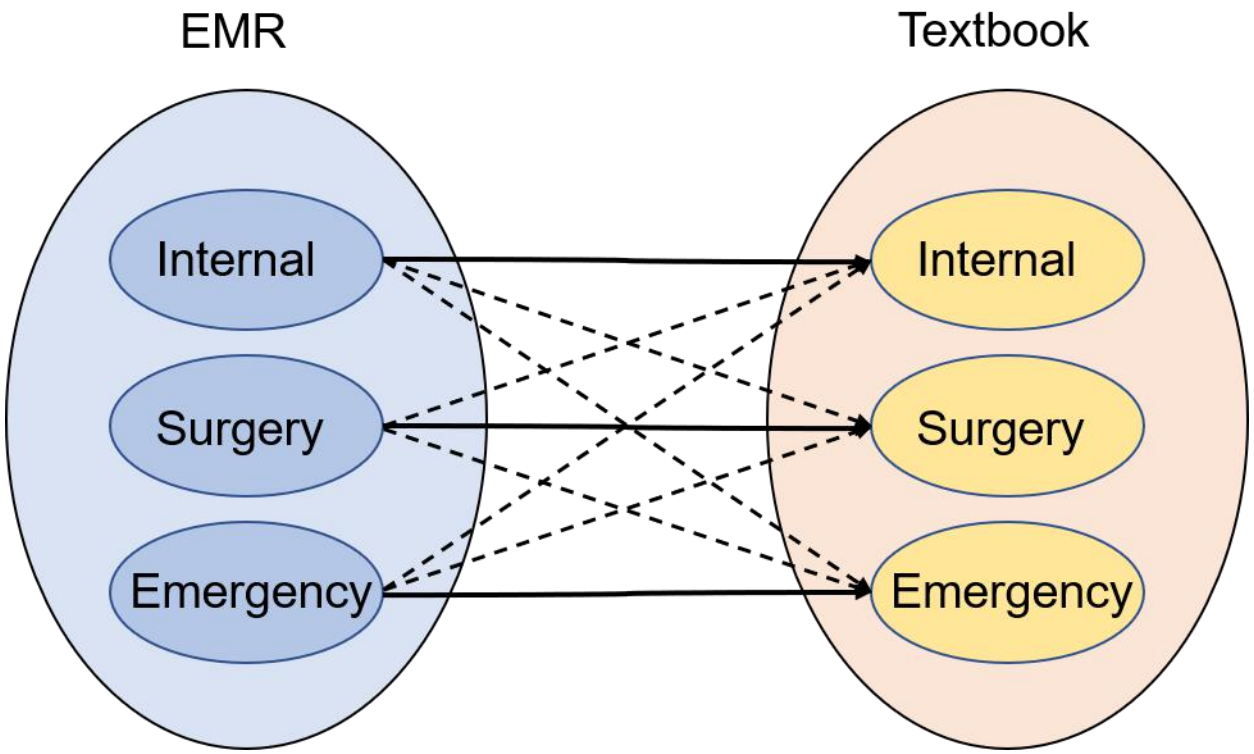
领域自适应标准范式

- **Single-Domain to Single-Domain adaptation**
- **Multi-Domain to Single-Domain adaptation**
- **Single-Domain to Multi-Domain adaptation**

背景

Internal Medicine	病人 病理 改变 显示 弥漫性肺泡损伤 (Disease) Patient's pathological changes show Diffuse alveolar damage, 和 炎症细胞浸润, (Disease) 早期的 特征 是 and inflammatory cell infiltration, the early feature is 肺水肿。 (Disease) pneumonedema.
Surgery	术前 心电图检查, (Check) 全麻下 Before surgery, ECG examination, under general anesthesia 行 动脉导管结扎术。 (Check) conduct ligation of patent ductus arteriosus.
Emergency	体检 发现 血压下降, (Sign) Physical examination reveals a drop in blood pressure, 膀胱占位, (Symptom) 排尿困难, (Symptom) 肉眼血尿。 (Symptom) bladder occupying, dysuria, gross hematuria.

Multi-domain to Multi-domain Adaptation

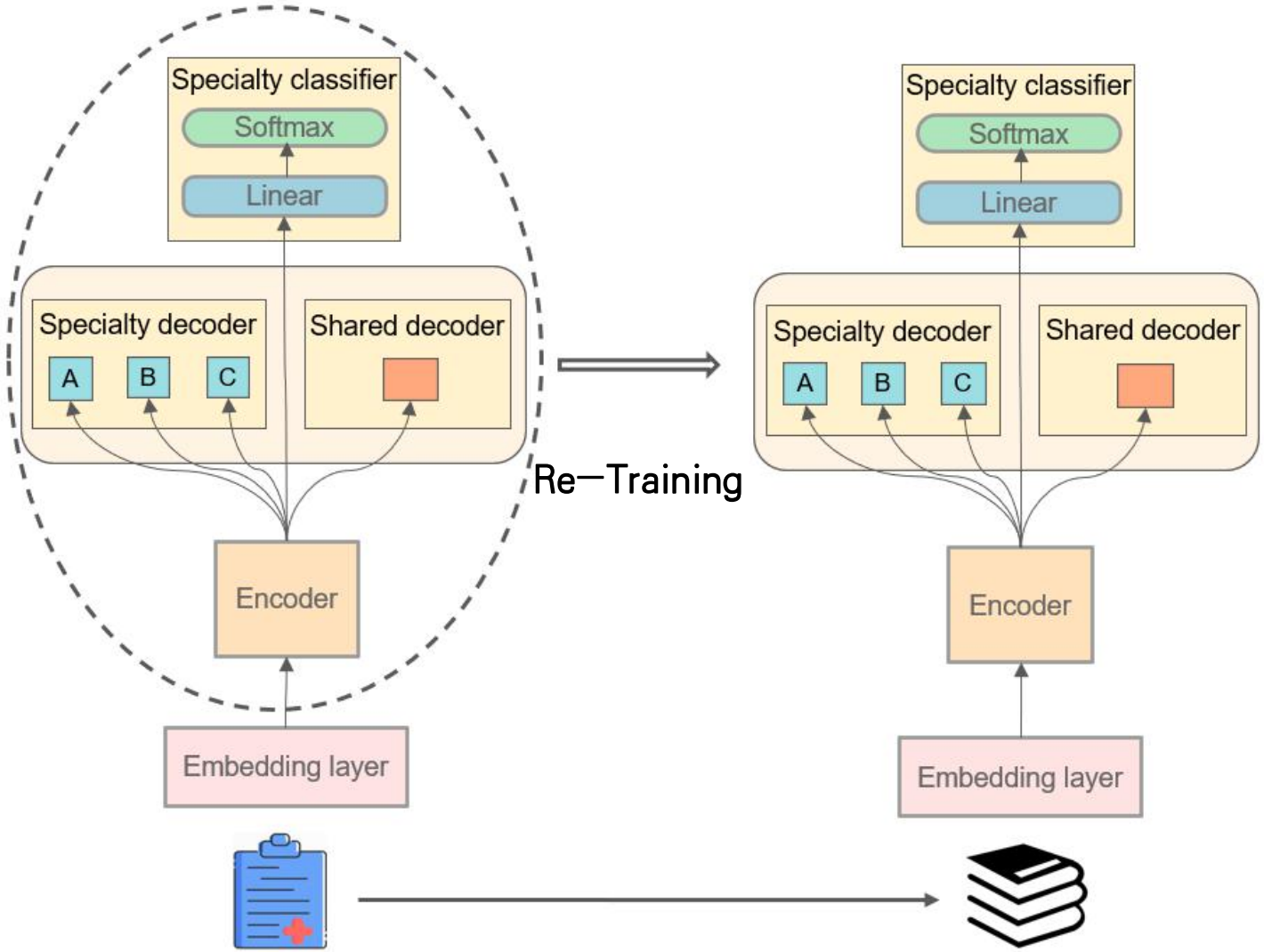


Specialty	EMR	Textbook
Internal Medicine	4521	492
Surgery	1782	18
Emergency	1856	227
Sum	8159	737

方法

Multi—Task ： MT

Multi—Specialty ： MS



实验设计

Without Domain adaptation （BBC/Target-Only、BBC/Mix、Lattice/Mix)

Single-Domain to Single-Domain adaptation **(**DANN、BBC/Re-Training)

Multi-Domain to Multi-Domain adaptation **(**BBC/Re-Training/MT、**MSADA**)

结果及消融实验

Method	Disease	Symptom	Sign	Operation	Drug	Check	F1-micro
BBC/Target-Only	83.94	76.24	59.01	22.00	83.72	55.13	73.92
BBC/Mix	87.31	83.12	45.05	47.33	73.44	61.82	76.17
Lattice/Mix [15]	**87.34**	81.68	50.79	64.24	75.49	72.78	78.51
DANN [16]	78.76	69.87	52.91	22.47	64.20	50.34	66.76
BBC/Re-Training	86.46	81.28	58.52	**72.29**	**85.61**	70.09	78.83
BBC/Re-Training/MT	86.86	82.92	**64.55**	45.67	83.52	69.06	79.23
MSADA	84.06	**86.51**	48.43	68.63	80.53	**80.11**	**80.84**

Method	Disease	Symptom	Sign	Operation	Drug	Check	F1-micro
MSADA	84.06	**86.51**	48.43	68.63	80.53	**80.11**	**80.84**
-MS	86.86	82.92	**64.55**	45.67	83.52	69.06	79.23
-MT	**87.28**	82.83	48.00	77.91	80.30	72.40	79.12
-BERT	82.00	75.35	27.51	47.20	53.94	56.13	68.64
-MT/MS	86.46	81.28	58.52	**72.29**	**85.61**	70.09	78.83
-MT/MS/BERT	78.34	72.64	48.23	20.41	57.93	54.76	66.25

总 结

In summary, this paper makes the following contributions:

1) We propose a novel multi-to-multi domain adaptation paradigm for Chinese MER, constructed a corresponding dataset, and explored the effect of various methods on the proposed dataset.

2) To solve the unique challenge of multi-to-multi domain adaptation from EMR to medical textbook, we propose a method named MSADA with adaptive specialty alignment mechanism to capture multi-specialty information.

3) On the dataset we constructed, MSADA significantly outperforms all other competitive methods, with improvements ranging between +1.61 to +14.08 F1-micro.

Multi-Specialty Domain Adaptation for Chinese Medical Named Entity Recognition

Thanks!