基于预训练语言模型的小样本医疗事件抽取系统

任务3-子任务2：面向中文电子病历的医疗事件抽取

戴松泰

Baidu 百度 知识图谱部

目 录

CONTENT

目 录

CONTENT

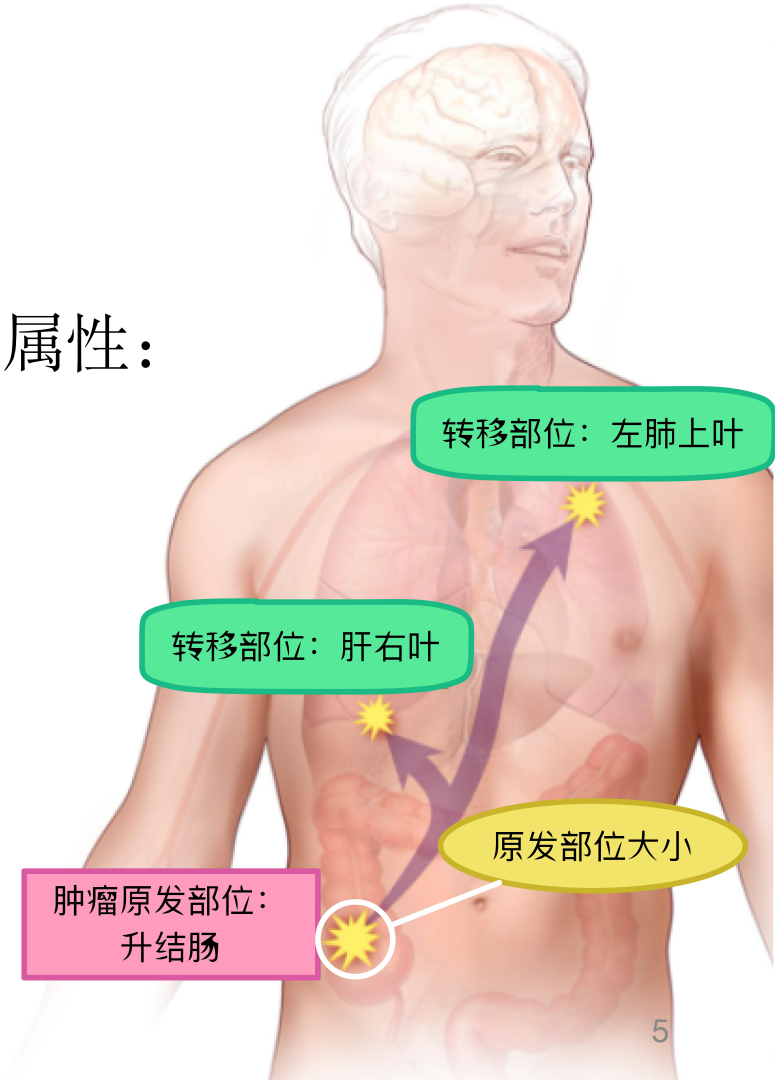
任务背景

- 任务类型：医疗文本知识化、结构化
- 难点：小样本
- 评测任务：面向中文电子病历的医疗事件抽取

任务定义

给定电子病历文本，抽取出肿瘤事件属性：

- 肿瘤原发部位
- 原发病灶大小
- 转移部位



任务示例

输入	右肺上叶周围型肺癌**双肺**多发转移瘤,纵隔淋巴结转移,双侧胸腔积液、心包积液,胸椎骨质破坏,较前(2015.7.15)进展。右肺上叶纵隔旁见椭圆形肿块影,大小约5.1*2.7CM,CT值19HU;两肺多发大小不等的结节影,大者直径约2.8CM。右肺上叶后段支气管截断。**双肺**门影增大。纵隔内多发增大的淋巴结,大者短径约2.0CM,其内见钙化灶。两胸腔见积液征象,右侧为著。冠脉钙化。心包积液。上胸椎骨质破坏,周围见软组织肿块影。
输出	肿瘤原发部位:右肺上叶 原发病灶大小:5.1CM×2.7CM 转移部位:双肺,纵隔淋巴结

任务难点

- 样本量少

任务数据

小样本场景

数据	数据量
训练集	1000
验证集	400
测试集	300
实体词表	863
未标注文本	1300

评价指标

基于属性实体的微平均 F1 值

目 录

CONTENT

答案位置回标

输入	右肺上叶周围型肺癌,**双肺**多发转移瘤,纵隔淋巴结转移,双侧胸腔积液、心包积液,胸椎骨质破坏,较前(2015.7.15)进展。右肺上叶纵隔旁见椭圆形肿块影,大小约5.1*2.7CM,CT值19HU;两肺多发大小不等的结节影,大者直径约2.8CM。右肺上叶后段支气管截断。**双肺**门影增大。纵隔内多发增大的淋巴结,大者短径约2.0CM,其内见钙化灶。两胸腔见积液征象,右侧为著。冠脉钙化。心包积液。上胸椎骨质破坏,周围见软组织肿块影。
输出	肿瘤原发部位：右肺上叶 原发病灶大小：5.1CM×2.7CM 转移部位：双肺,纵隔淋巴结

原始数据没有答案位置

答案位置回标

输入	右肺上叶周围型肺癌,双肺多发转移瘤,纵隔淋巴结转移,双侧胸腔积液、心包积液,胸椎骨质破坏,较前(2015.7.15)进展。右肺上叶纵隔旁见椭圆形肿块影,大小约5.1*2.7CM,CT值19HU;两肺多发大小不等的结节影,大者直径约2.8CM。右肺上叶后段支气管截断。双肺门影增大。纵隔内多发增大的淋巴结,大者短径约2.0CM,其内见钙化灶。两胸腔见积液征象,右侧为著。冠脉钙化。心包积液。上胸椎骨质破坏,周围见软组织肿块影。
输出	肿瘤原发部位：右肺上叶 原发病灶大小：5.1CM×2.7CM 转移部位：双肺,纵隔淋巴结

字面匹配?

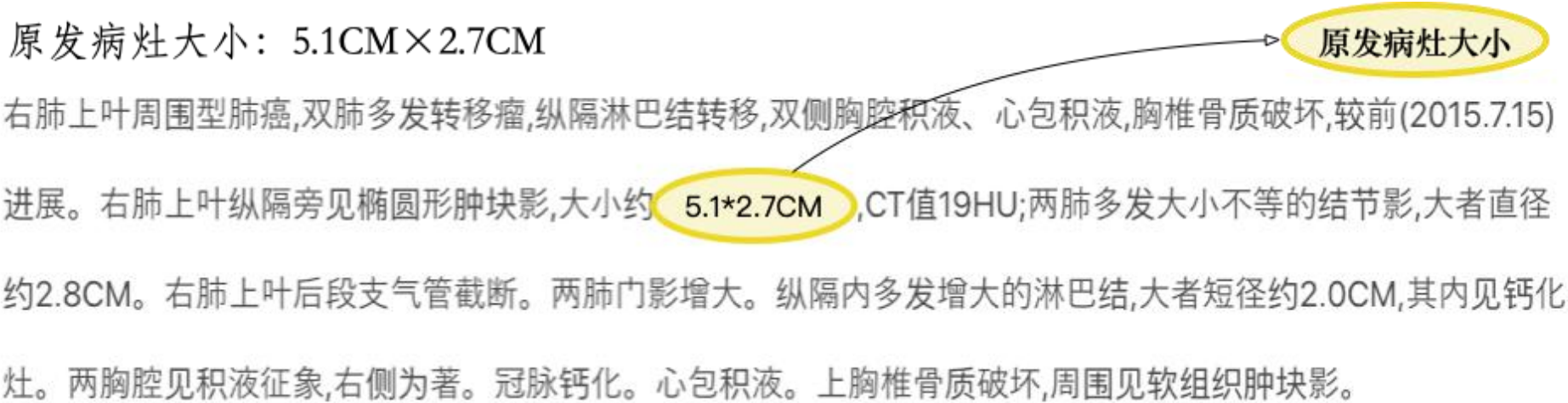
答案位置回标

输入	右肺上叶周围型肺癌,**双肺**多发转移瘤,纵隔淋巴结转移,双侧胸腔积液、心包积液,胸椎骨质破坏,较前(2015.7.15)进展。右肺上叶纵隔旁见椭圆形肿块影,大小约5.1*2.7CM,CT值19HU;两肺多发大小不等的结节影,大者直径约2.8CM。右肺上叶后段支气管截断。**双肺**门影增大。纵隔内多发增大的淋巴结,大者短径约2.0CM,其内见钙化灶。两胸腔见积液征象,右侧为著。冠脉钙化。心包积液。上胸椎骨质破坏,周围见软组织肿块影。
输出	肿瘤原发部位：右肺上叶 原发病灶大小：5.1CM×2.7CM 转移部位：双肺,纵隔淋巴结

字面匹配?

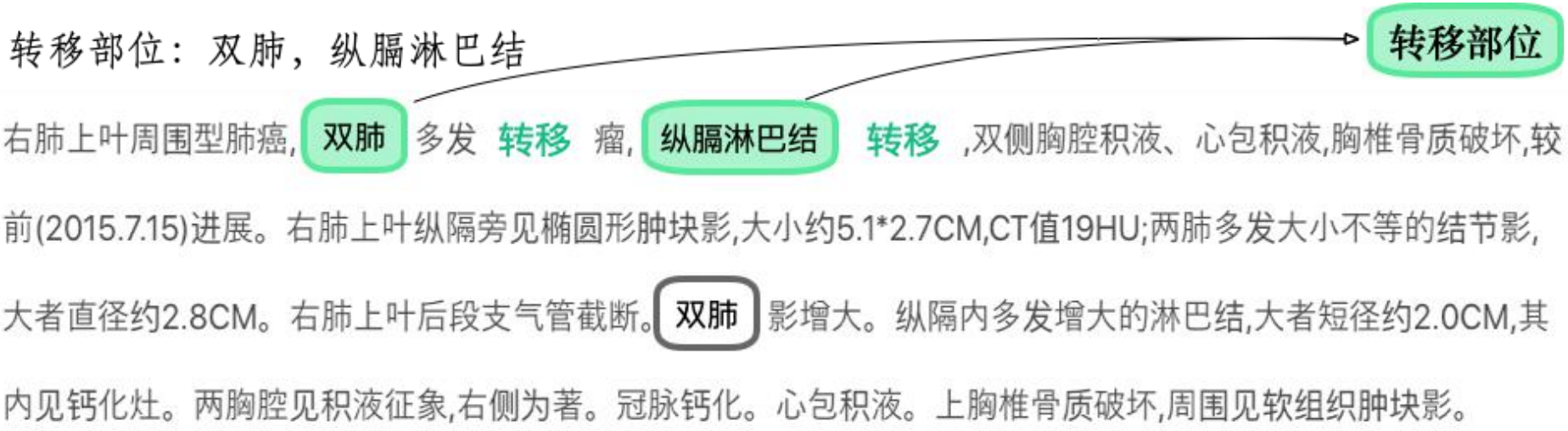
答案位置回标：原发病灶大小

- 根据正则表达式标注



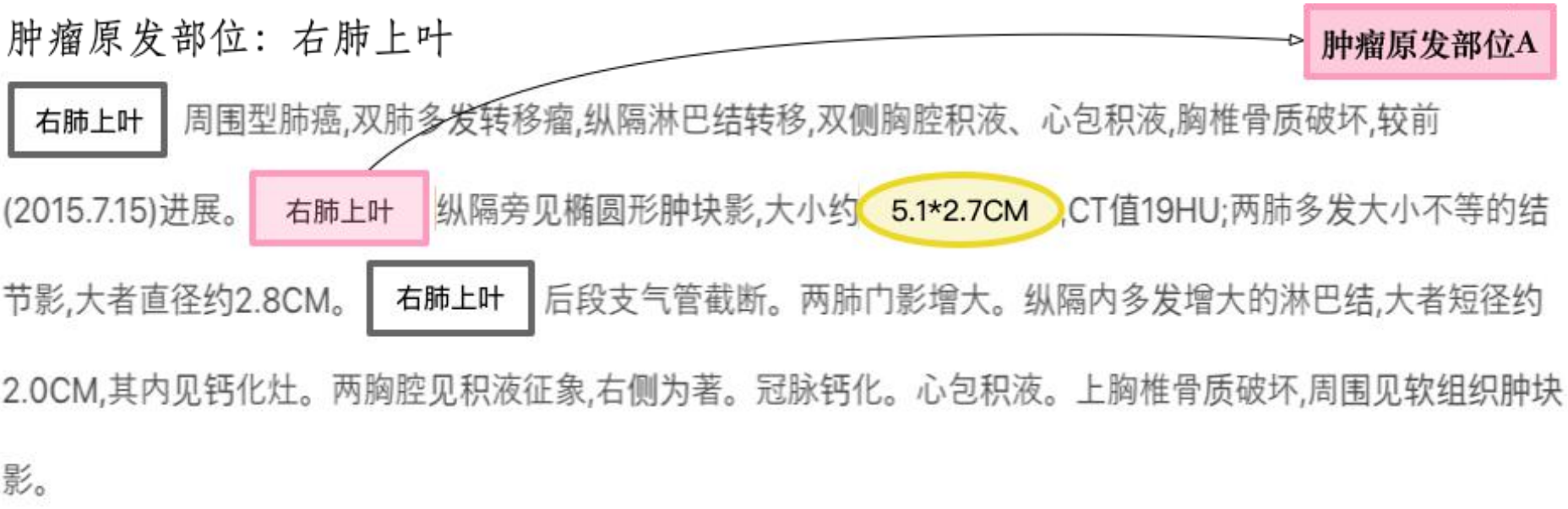
答案位置回标：转移部位

- 保留位于“转移”附近的部位



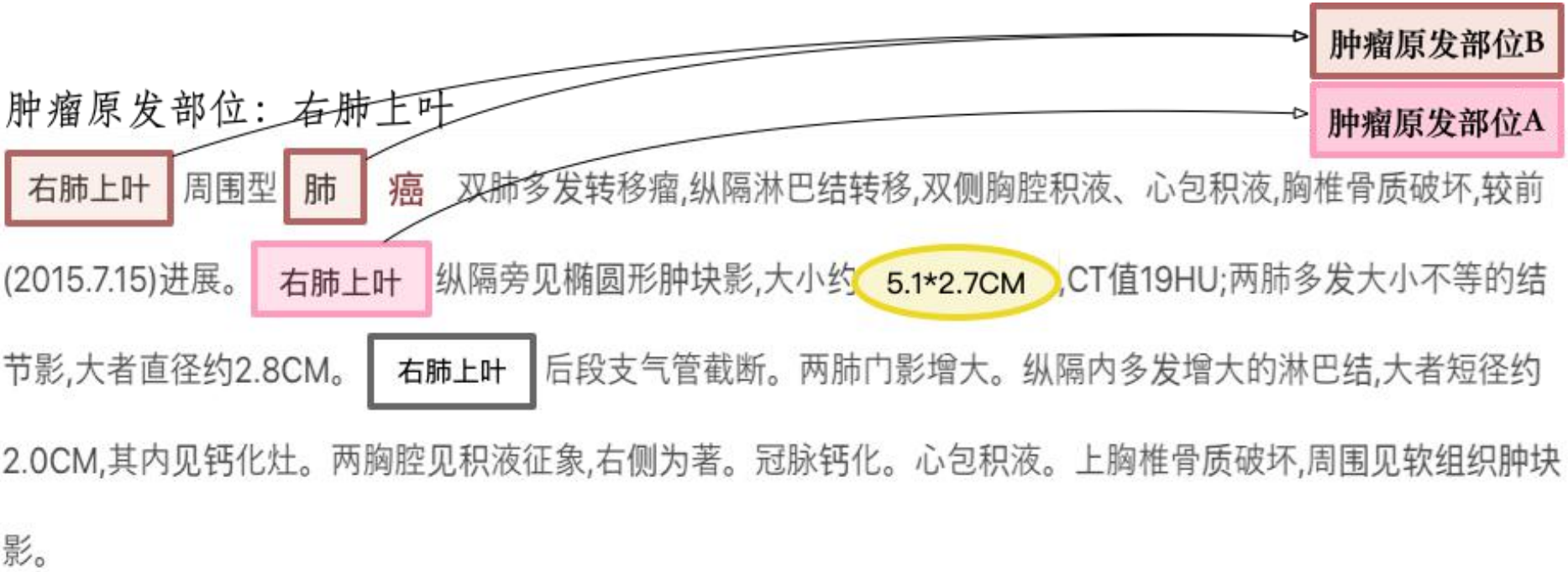
答案位置回标：肿瘤原发部位

- 一种情况：保留位于「原发病灶大小」附近的部位



肿瘤原发部位回标

- 另一种情况：保留位于“癌”，“恶性”等词附近的部位



属性槽位

原发病灶大小

转移部位

肿瘤原发部位A

肿瘤原发部位B

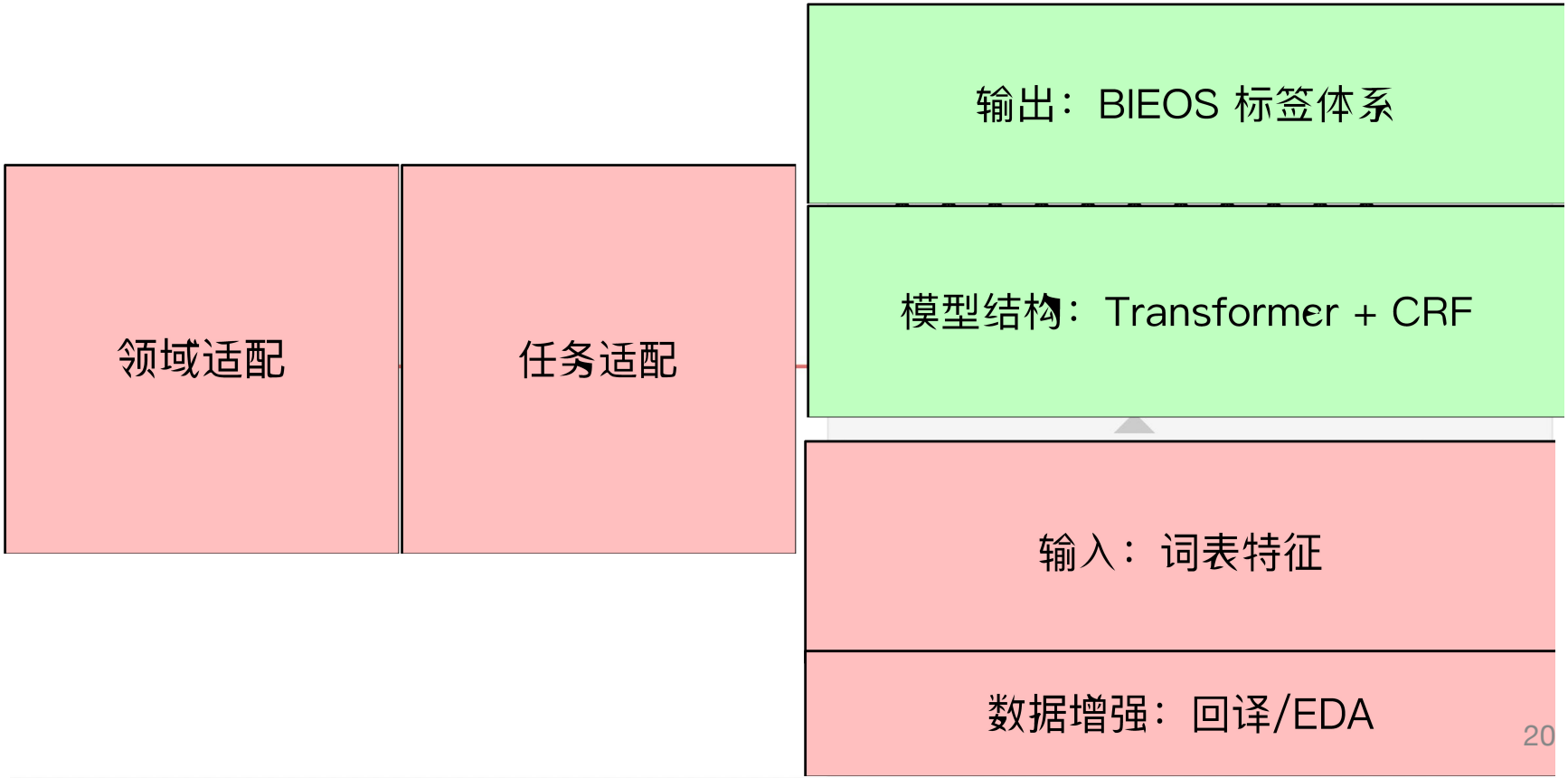
文本规范化

癌症表述	训练集频率	测试集频率
癌	690	166
MT	96	0
CA	1	87
cancer	0	42
carcinoma	0	4

目 录

CONTENT

模型训练流程

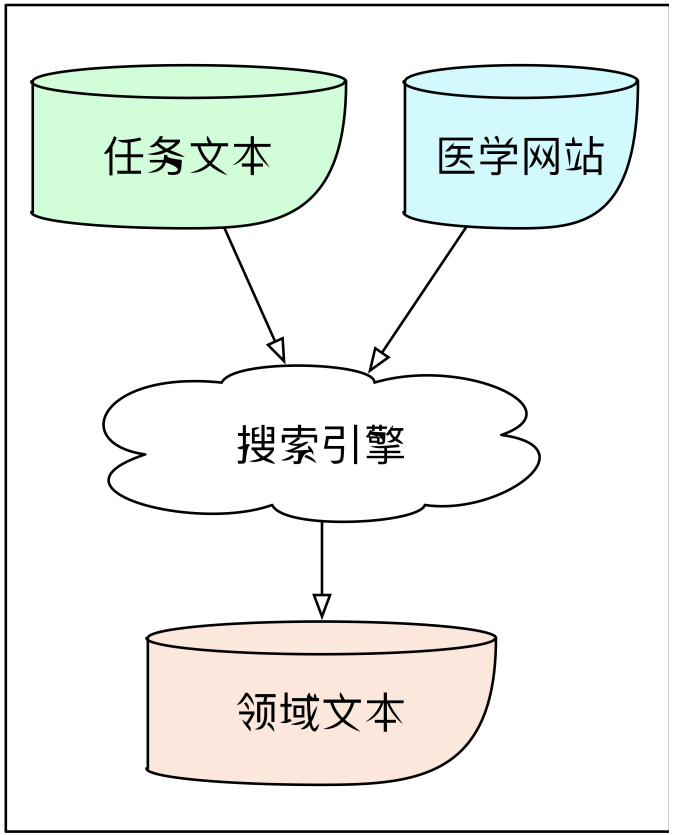

输出：BIEOS 标签体系
模型结构：Transformer + CRF
输入：词表特征
数据增强：回译/EDA
领域适配
任务适配

领域适配 & 任务适配

- 语言模型：海量通用语料预训练
- 电子病历：特定领域，特定文本风格
- 样本量小：对语言建模能力提升有限

- 领域适配：领域语料预训练
- 任务适配：任务文本预训练

领域适配：语料获取

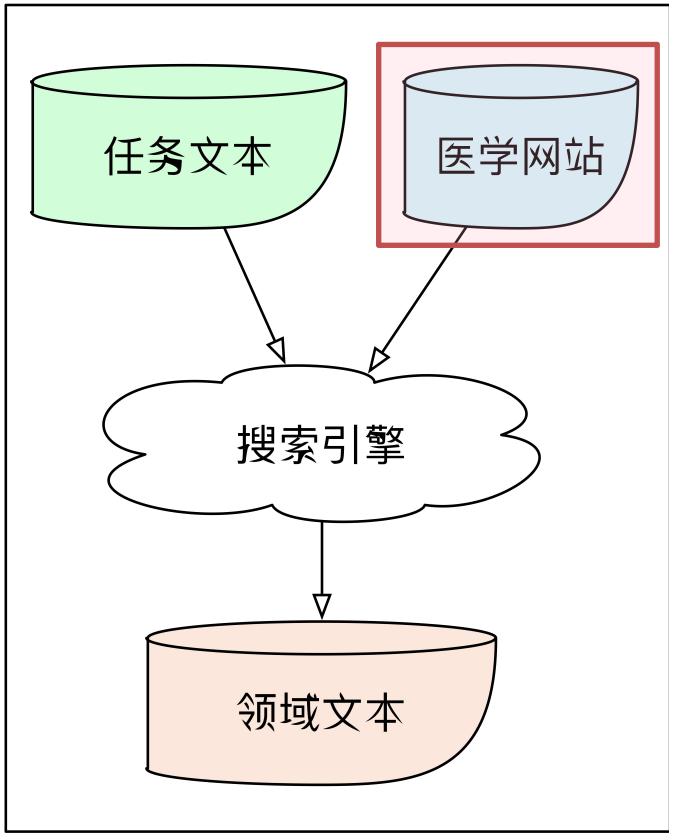


与肿瘤诊断相关的医学文本

- 医学书籍：版权，数据量小
- 电子病历：隐私
- 网络语料：分布零散

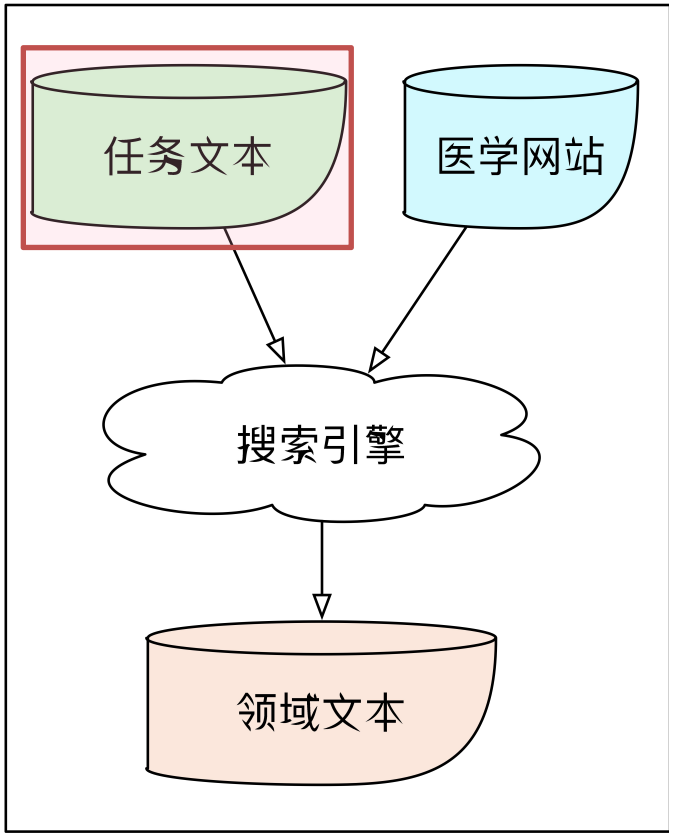
- **搜索引擎**

领域适配：语料获取



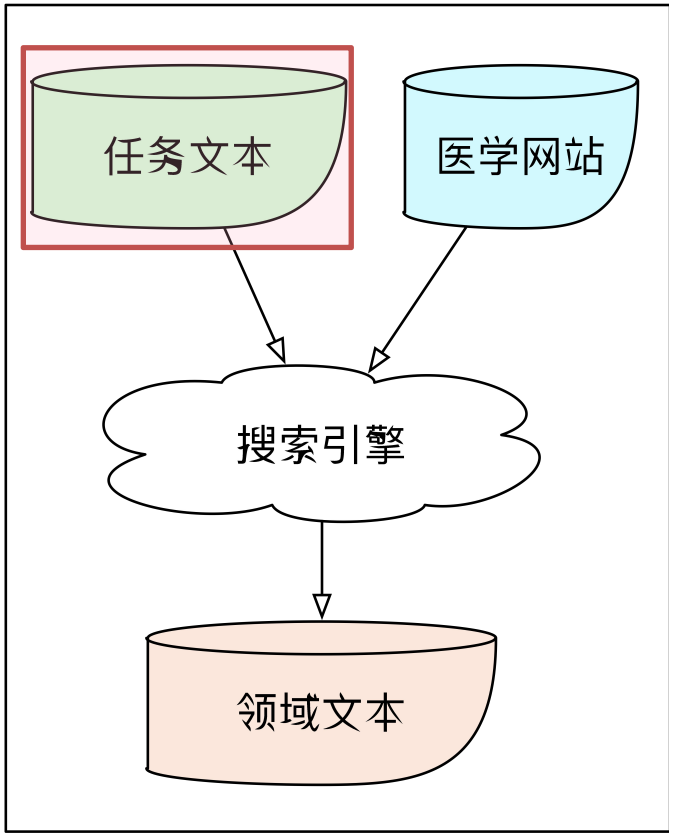
专业网站	网址
影像PPT	yxppt.com
寻医问药	xywy.com
快速问医生	120ask.com
好大夫	haodf.com
丁香园	dxy.cn
医联	medlinker.com
医脉通	medlive.cn
健康界	cn-healthcare.com
天山医学院	tsu.tw
中国医学会核医学分会	chinanm.cma.org.cn
医生在线	51daifu.com

领域适配：语料获取



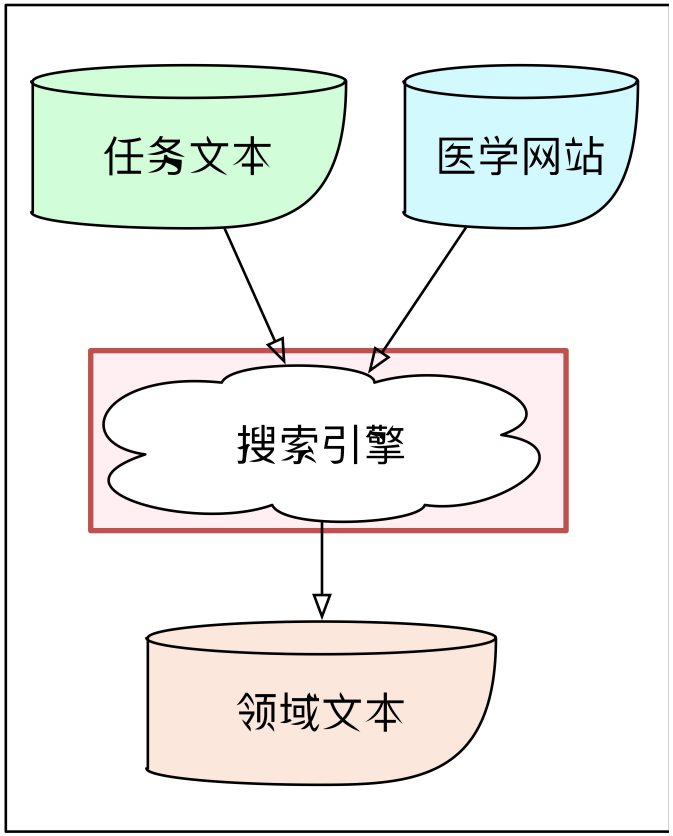
肝右叶下段可见类圆形混杂信号影，其内可见不规则长t1长t2信号区，增强扫描动脉期病灶实性成份明显不均匀强化，内可见迂曲小血管影，后强化程度下降

领域适配：语料获取



肝右叶下段可见类圆形　混杂信号影，其内可见不规则长t1　长t2信号区，增强扫描动脉期病灶实性　成份明显不均匀强化，内可见迂曲小血管影，后强化程度下降

领域适配：语料获取



增强扫描动脉期病灶实性 site:yxppt.com

网页 资讯 视频 图片 知道 文库 贴吧 地图

时间不限 所有网页和文件 yxppt.com 清除

【病例】肝实性间叶性错构瘤1例CT影像表现解析 – 影像PPT

2017年4月15日 3.CT平扫肝脏类圆形实性稍低密度肿块,边界清晰,肿块占据肝右叶和左叶内侧段,胆管及肝内门静脉分支受压移位。4.增强扫描动脉期病灶内云絮状轻微强化,包...

www.yxppt.com/archives/216... 百度快照

【病例】肝肉瘤样癌1例CT – 影像PPT

2017年3月26日 (a):CT平扫示肝右叶低密度占位,约5.0*5.3cm,边界不清,病灶中央见片状囊性密度区 (b):CT增强扫描动脉期,病灶边缘实性成分轻度强化 ...

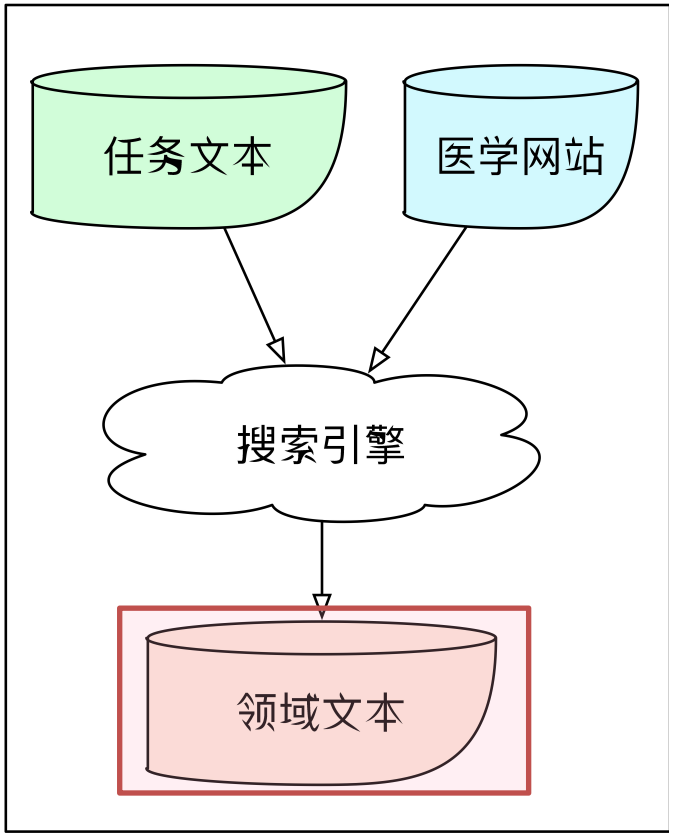
www.yxppt.com/archives/210... 百度快照

【病例】肝脏淋巴上皮瘤样癌三例 – 影像PPT

2016年12月30日 CT增强扫描动脉期示病灶实性成分明显不均匀强化(图7)。CT增强扫描门静脉期示病灶实性成分与肝实质强化程度相似,其内囊变坏死区未见明确强化(图8)...

www.yxppt.com/?p=188... 百度快照

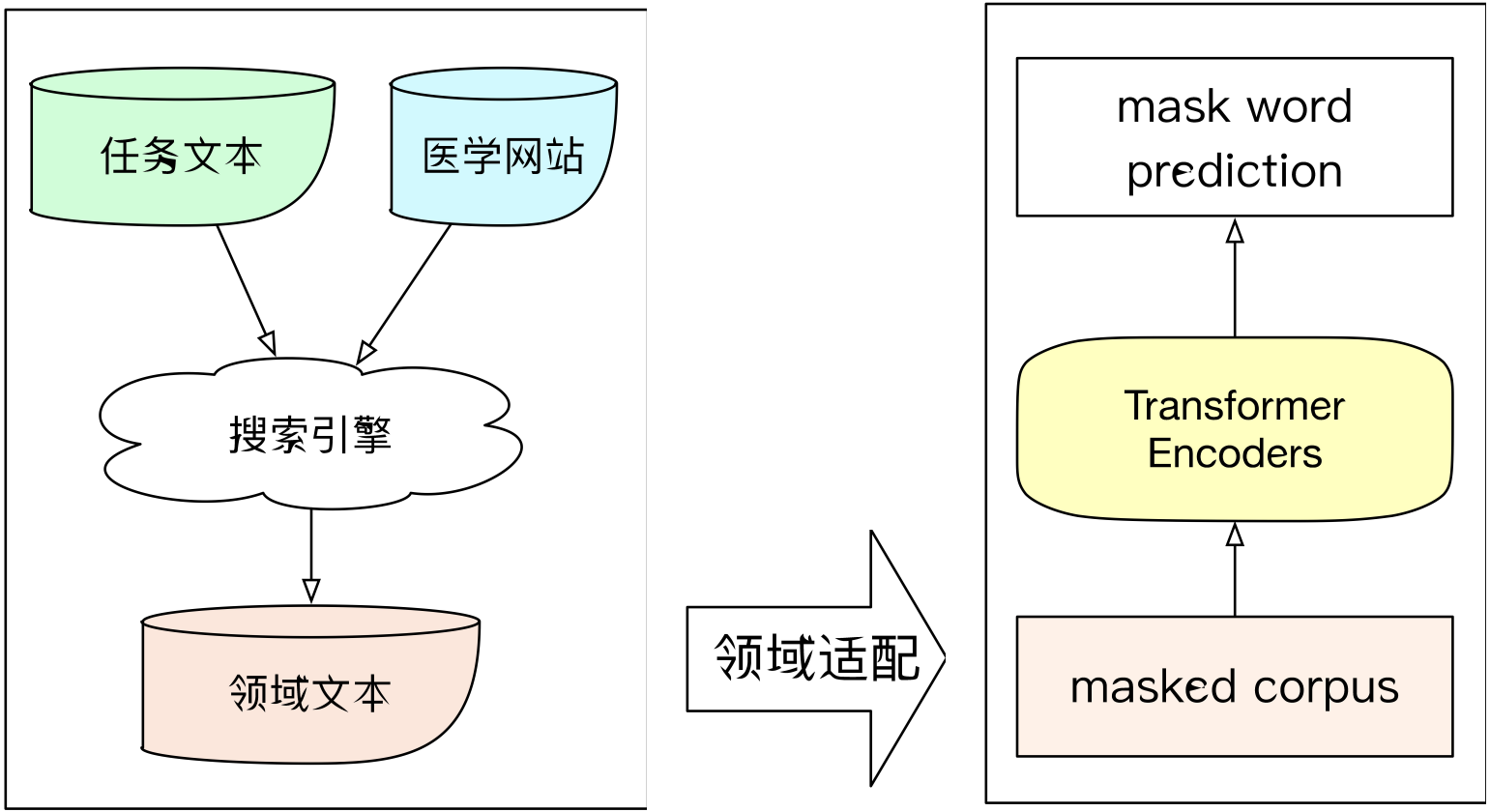
领域适配：语料获取



Query:增强扫描动脉期病灶实性 site:yxppt.com

领域文本：CT平扫肝右叶巨块状低密度灶，边缘欠规整，密度不均匀，肿块局部肝脏轮廓膨隆伴有小范围凹陷；增强扫描动脉期病灶内不规则斑片状及类似血池状强化，边缘和中心部位不强化；门脉期强化部位持续强化并向外扩散，强化范围有扩大，密度略高于肝实质密度，中心区域不强化；平衡期病灶强化部分密度稍高于肝脏实质密度，与正常肝实质分界不清，不强化区范围扩大，形态不光整；

领域适配


任务文本
医学网站
搜索引擎
领域文本
领域适配
mask word prediction
Transformer Encoders
masked corpus

对比：领域语料与任务文本

领域语料

肝间叶性错构瘤（hepatic mesenchyma hamartoma）为肝内少见的良性病变，在肝良性肿瘤中仅次于肝血管瘤。组织来源不清，大多数学者认为它是一种发育畸形或异常反应的结果，也有学者认为肝脏间叶性错构瘤与“19q13.4位点”断裂有关。通常5岁前发病，多见于4个月～2岁的婴幼儿，男性略多于女性。病理上分实性、囊性，数量上有单发和多发之分。

非电子病历

任务文本

肝脏体积缩小，表面不光滑，信号不均，呈结节样改变，各叶比例失调，左叶增大，肝右叶下段可见类圆形混杂信号影，主要呈稍长t1稍长t2信号，大小约67.2*80.8mm，其内可见不规则长t1长t2信号区，增强扫描动脉期病灶实性成份明显不均匀强化，内可见迂曲小血管影，后强化程度下降，延迟期低于周围正常肝实质信号...　电子病历

对比：领域语料与任务文本

领域语料

大部分肿瘤诊断相关

少部分是电子病历

肝间叶性错构瘤（hepatic mesenchyma hamartoma）为肝内少见的良性病变，在肝良性肿瘤中…清，大多数学者认为它是一种发育畸形或异常反应的结果，也…瘤与“19q13.4位点”断裂有关。通常5岁前发病，多见于4个月～2岁…，病理上分实性、囊性，数量上有单发和多发之分。

非电子病历

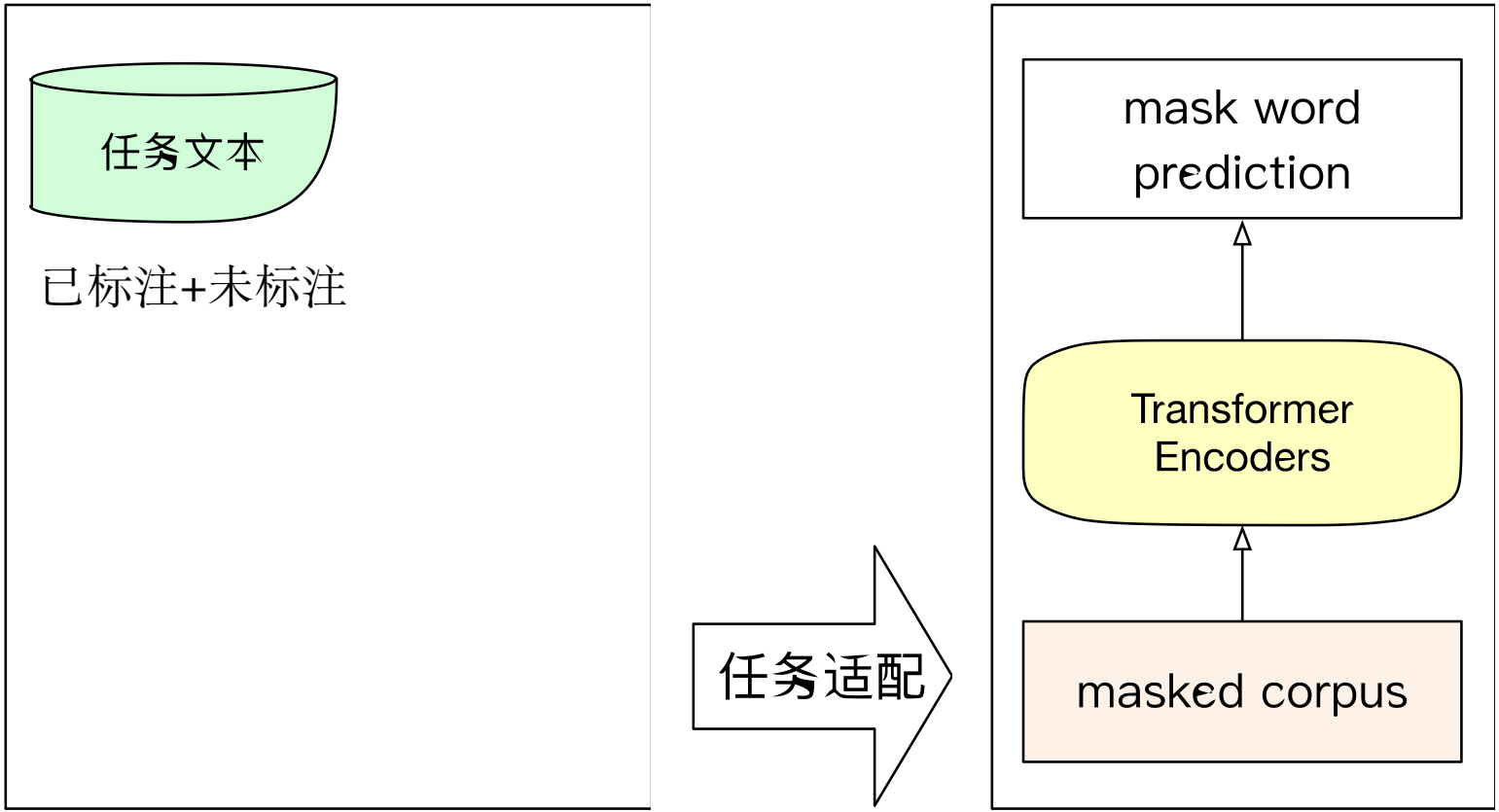
任务文本

全部是电子病历

肝脏体…号不均，呈结节样改变，各叶比例失调，左叶增大，肝右叶…影，主要呈稍长t1稍长t2信号，大小约67.2*80.8mm，其内可见不规则长t1长t2信号区，增强扫描动脉期病灶实性成份明显不均匀强化，内可见迂曲小血管影，后强化程度下降，延迟期低于周围正常肝实质信号…

电子病历

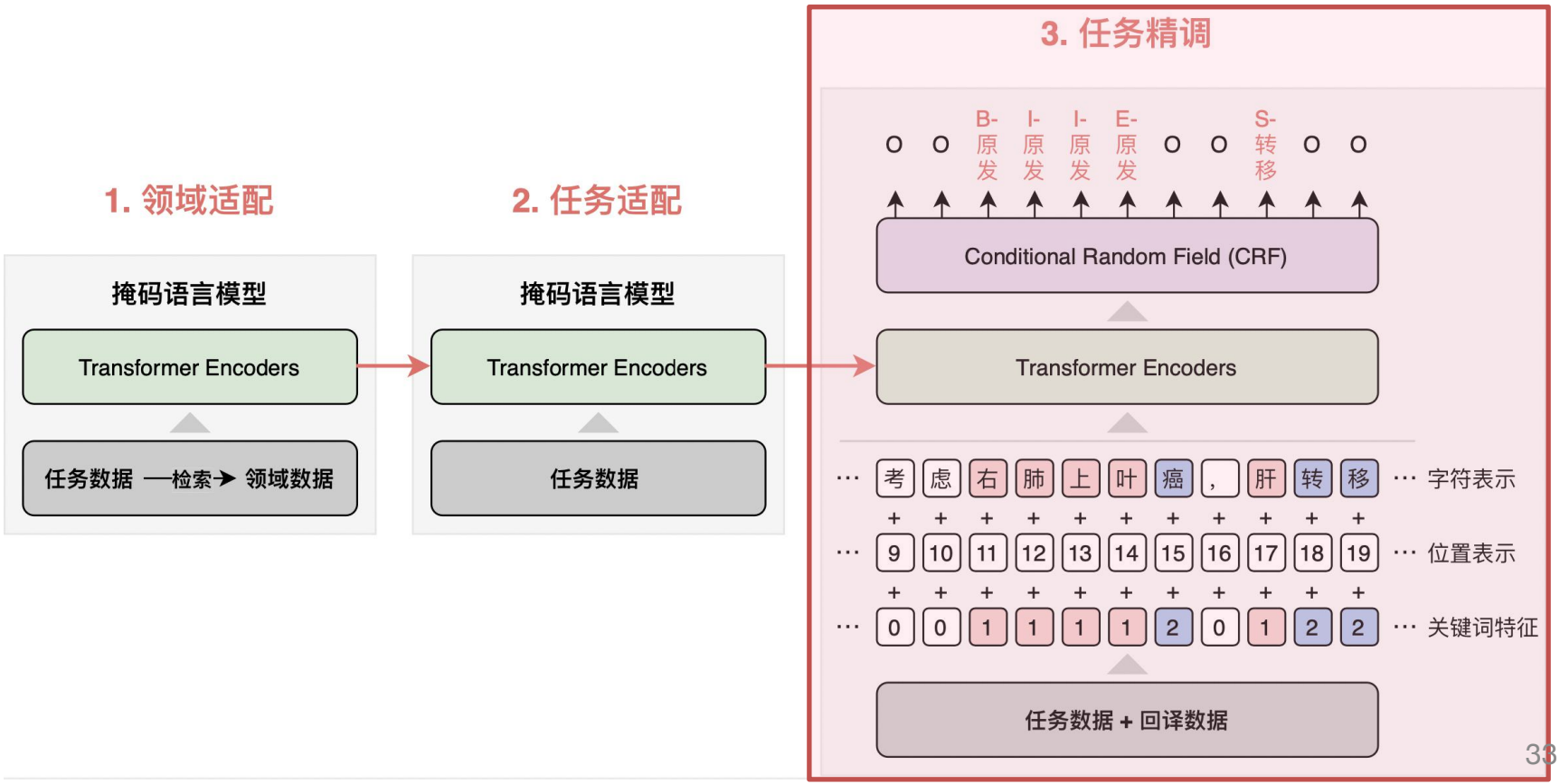
任务适配


任务文本
已标注+未标注
任务适配
mask word prediction
Transformer Encoders
masked corpus

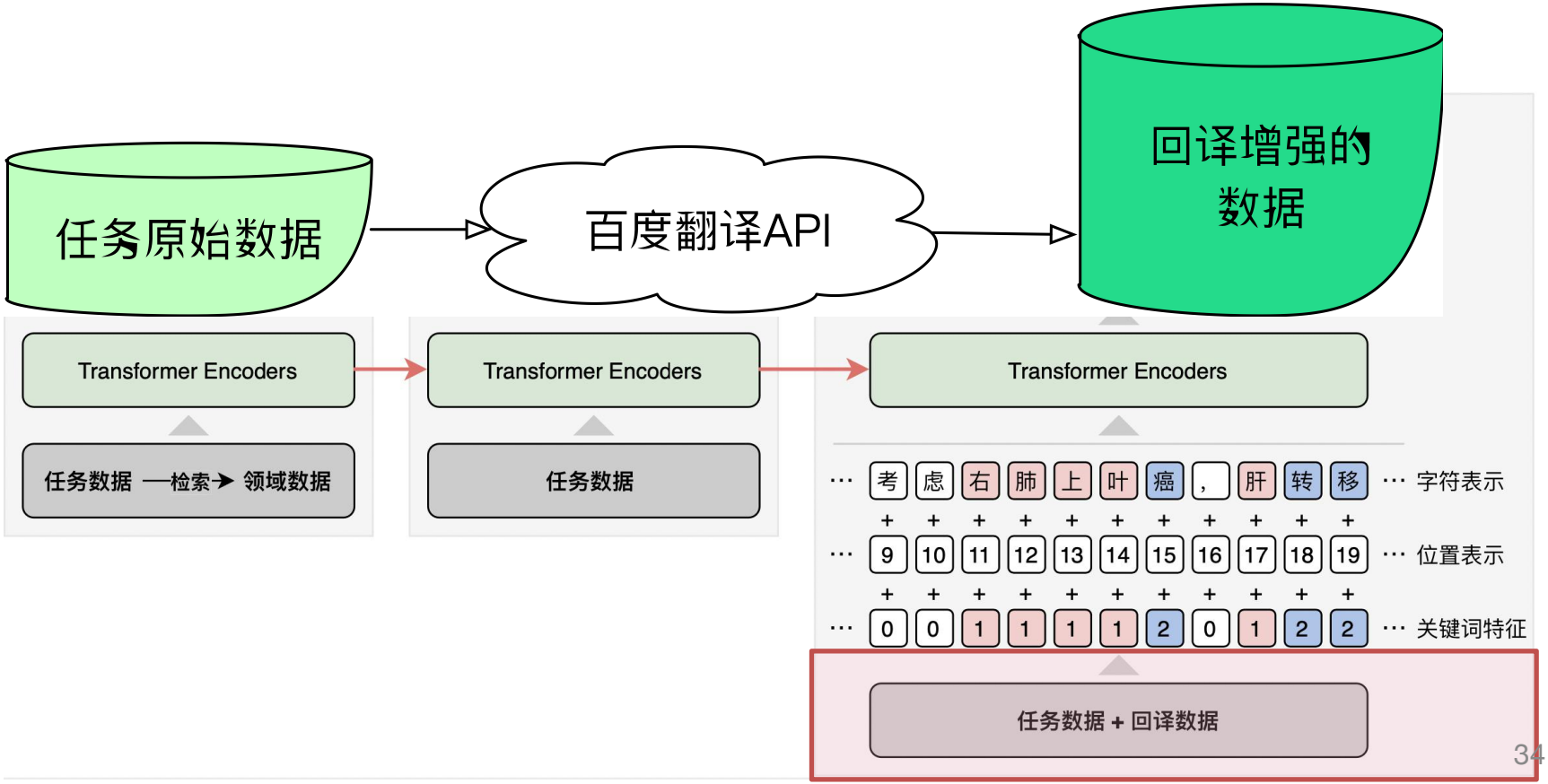
领域适配 & 任务适配

指标	领域适配	任务适配
语料字符数	2.6亿	134万
预训练轮数	10	100
训练时间 (8卡 V100)	50小时	4小时

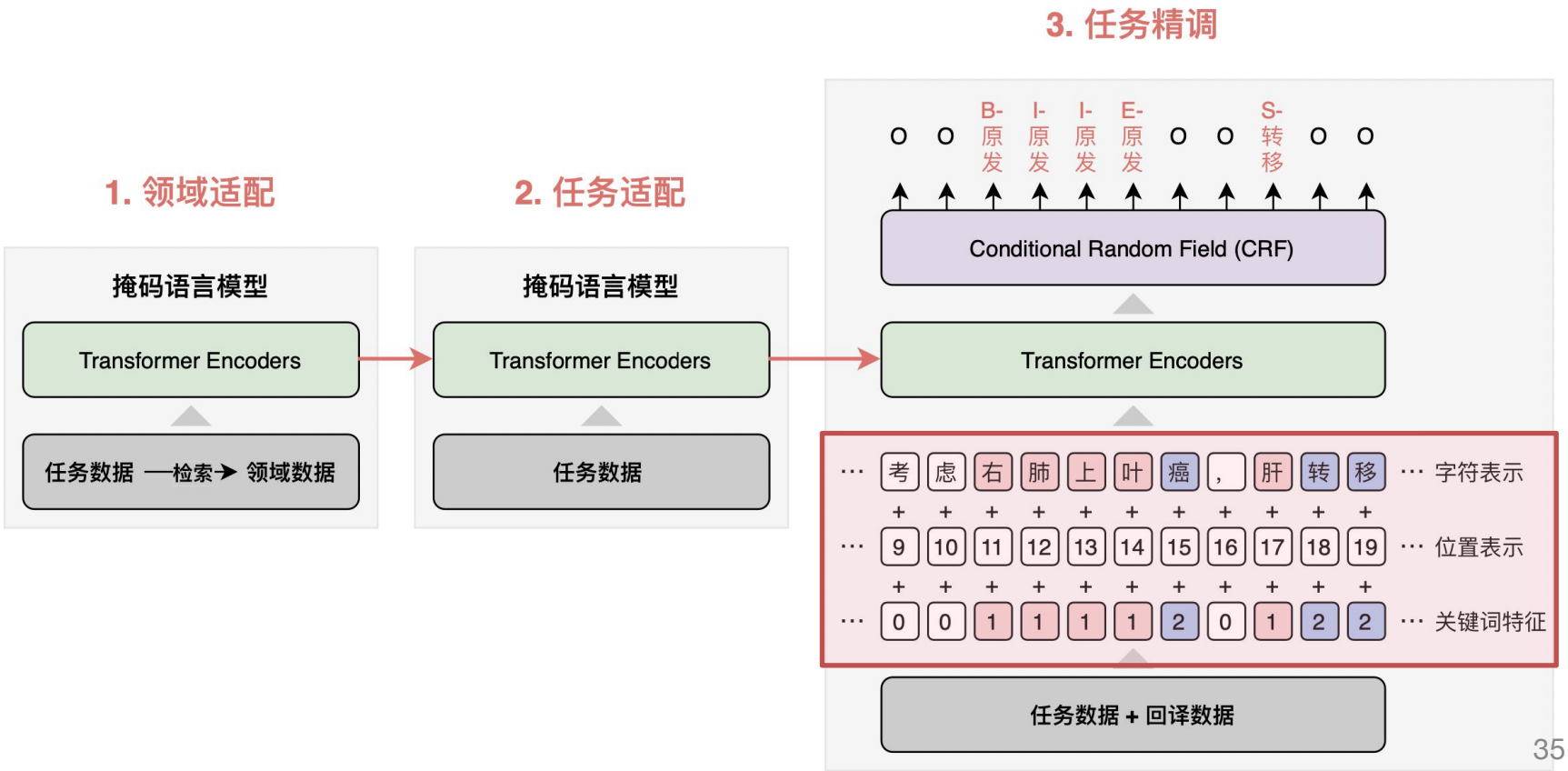
模型精调


1. 领域适配
掩码语言模型
Transformer Encoders
任务数据 —检索➔ 领域数据
2. 任务适配
掩码语言模型
Transformer Encoders
任务数据
3. 任务精调
O O B-原发 I-原发 I-原发 E-原发 O O S-转移 O O
Conditional Random Field (CRF)
Transformer Encoders
… 考 虑 右 肺 上 叶 癌 , 肝 转 移 … 字符表示
… 9 10 11 12 13 14 15 16 17 18 19 … 位置表示
… 0 0 1 1 1 1 2 0 1 2 2 … 关键词特征
任务数据 + 回译数据

数据增强：回译


任务原始数据
百度翻译API
回译增强的数据
Transformer Encoders
Transformer Encoders
Transformer Encoders
任务数据 —检索➔ 领域数据
任务数据
… 考 虑 右 肺 上 叶 癌 , 肝 转 移 … 字符表示
… 9 10 11 12 13 14 15 16 17 18 19 … 位置表示
… 0 0 1 1 1 1 2 0 1 2 2 … 关键词特征
任务数据 + 回译数据

模型输入：加入关键词


1. 领域适配
掩码语言模型
Transformer Encoders
任务数据 —检索➔ 领域数据
2. 任务适配
掩码语言模型
Transformer Encoders
任务数据
3. 任务精调
O O B-原发 I-原发 I-原发 E-原发 O O S-转移 O O
Conditional Random Field (CRF)
Transformer Encoders
… 考 虑 右 肺 上 叶 癌 , 肝 转 移 … 字符表示
… 9 10 11 12 13 14 15 16 17 18 19 … 位置表示
… 0 0 1 1 1 1 2 0 1 2 2 … 关键词特征
任务数据 + 回译数据

模型输入：加入关键词

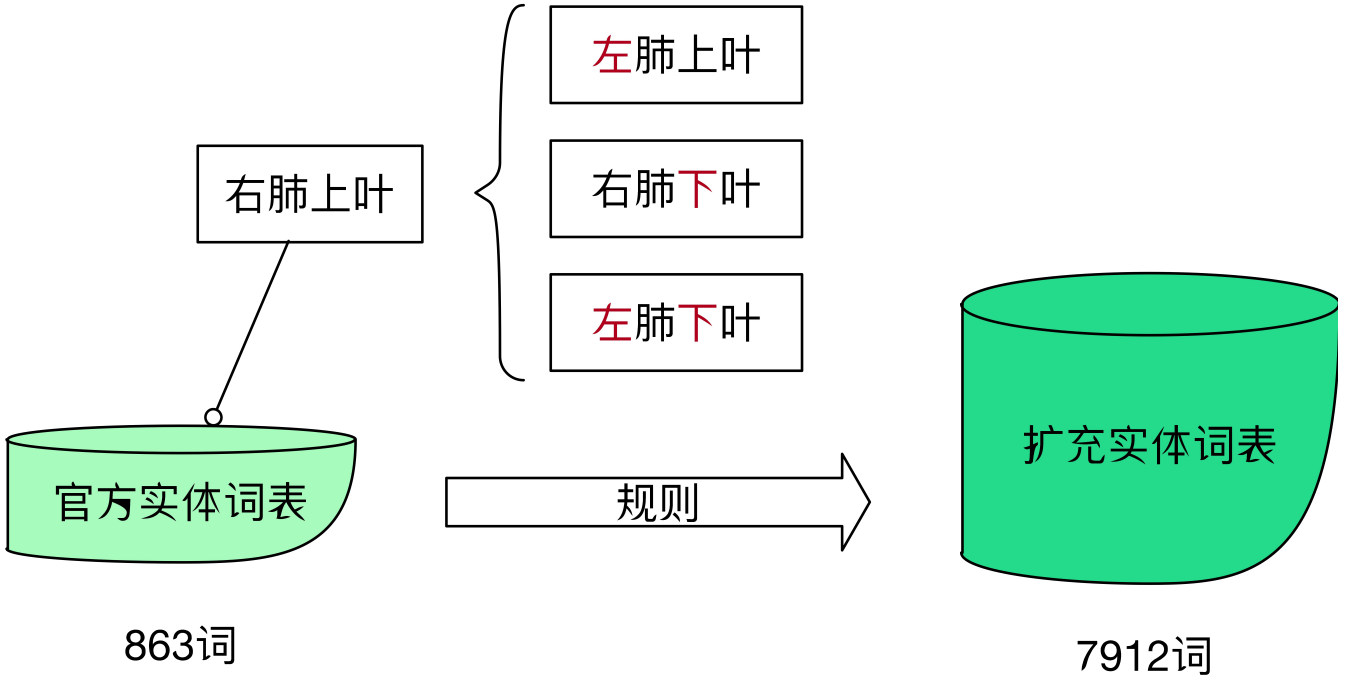
关键词
- 1. 实体词 （双肺，纵膈，淋巴结，右肺上叶）
- 2.答案提示词 （转移，癌）

转移部位：双肺，纵膈淋巴结

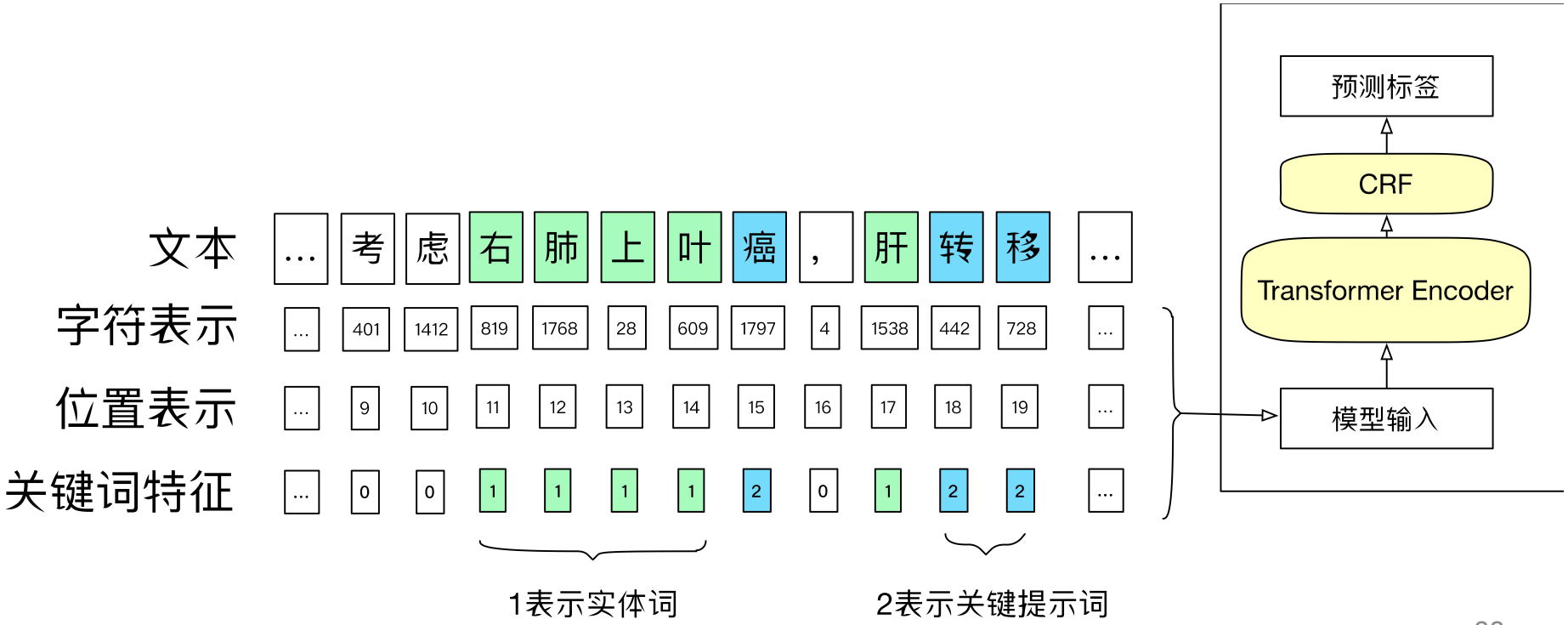
转移部位

右肺上叶周围型肺癌，双肺 多发 转移 瘤，纵膈淋巴结 转移 ,双侧胸腔积液、心包积液,胸椎骨质破坏,较前(2015.7.15)进展。右肺上叶纵隔旁见椭圆形肿块影,大小约5.1*2.7CM,CT值19HU;两肺多发大小不等的结节影,大者直径约2.8CM。右肺上叶后段支气管截断。双肺 影增大。纵隔内多发增大的淋巴结,大者短径约2.0CM,其内见钙化灶。两胸腔见积液征象,右侧为著。冠脉钙化。心包积液。上胸椎骨质破坏,周围见软组织肿块影。

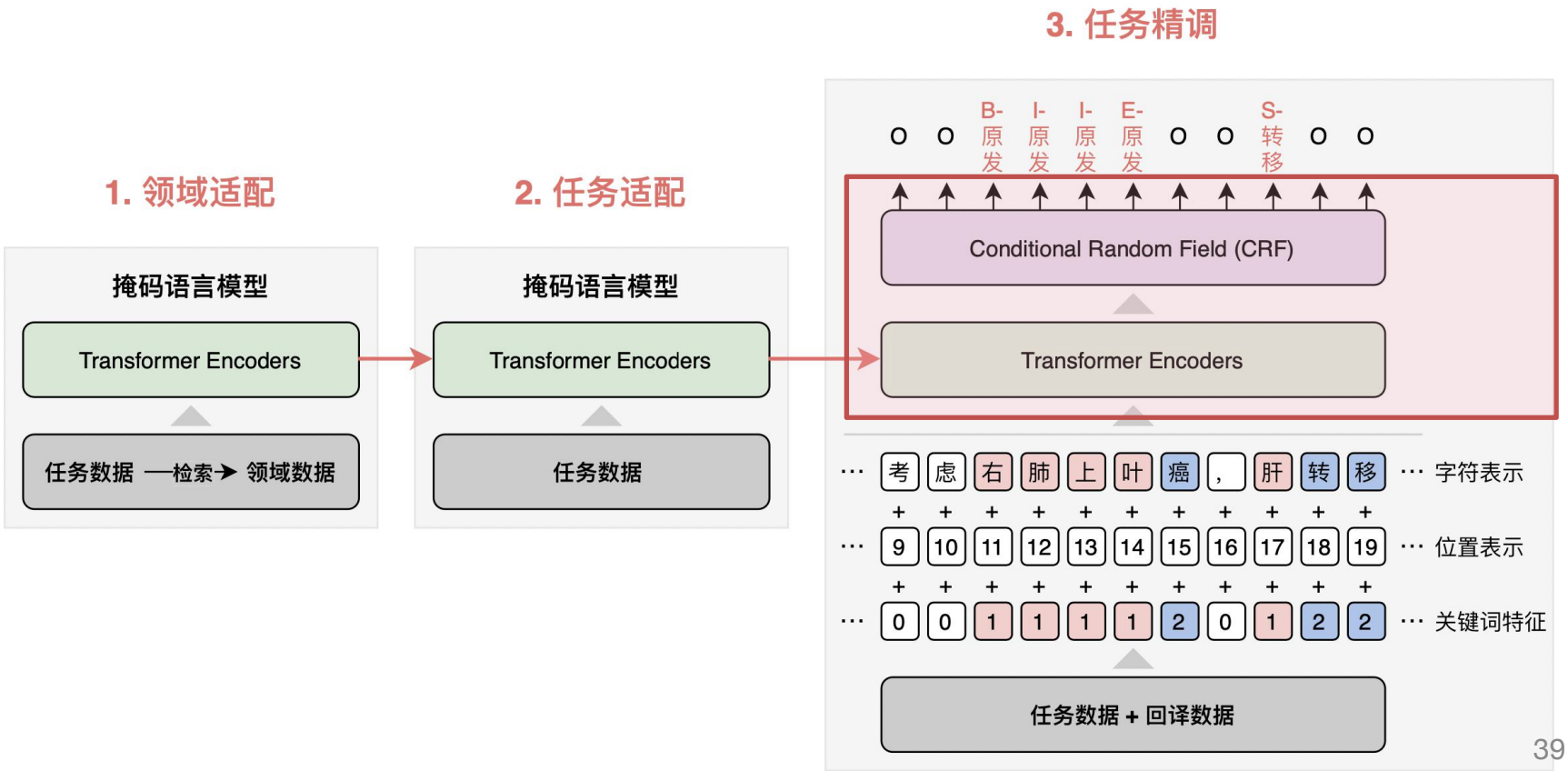
模型输入：加入关键词



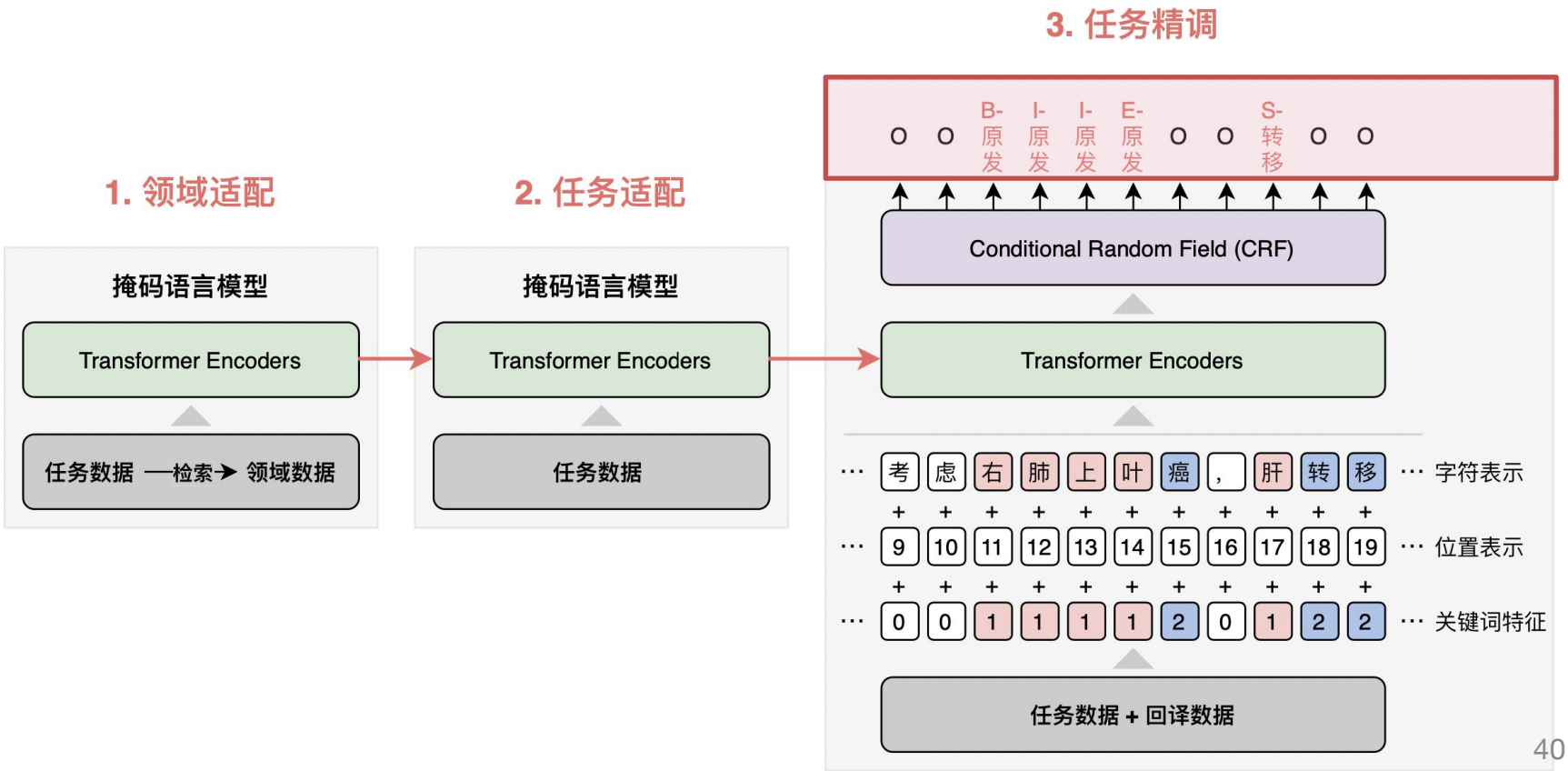
模型输入：加入关键词


文本
字符表示
位置表示
关键词特征
… 考 虑 右 肺 上 叶 癌 , 肝 转 移 …
… 401 1412 819 1768 28 609 1797 4 1538 442 728 …
… 9 10 11 12 13 14 15 16 17 18 19 …
… 0 0 1 1 1 1 2 0 1 2 2 …
1表示实体词
2表示关键提示词
预测标签
CRF
Transformer Encoder
模型输入

模型结构


1. 领域适配
掩码语言模型
Transformer Encoders
任务数据 —检索➔ 领域数据
2. 任务适配
掩码语言模型
Transformer Encoders
任务数据
3. 任务精调
O O B-原发 I-原发 I-原发 E-原发 O O S-转移 O O
Conditional Random Field (CRF)
Transformer Encoders
… 考 虑 右 肺 上 叶 癌 , 肝 转 移 … 字符表示
… 9 10 11 12 13 14 15 16 17 18 19 … 位置表示
… 0 0 1 1 1 1 2 0 1 2 2 … 关键词特征
任务数据 + 回译数据

模型输出：BIEOS标签体系

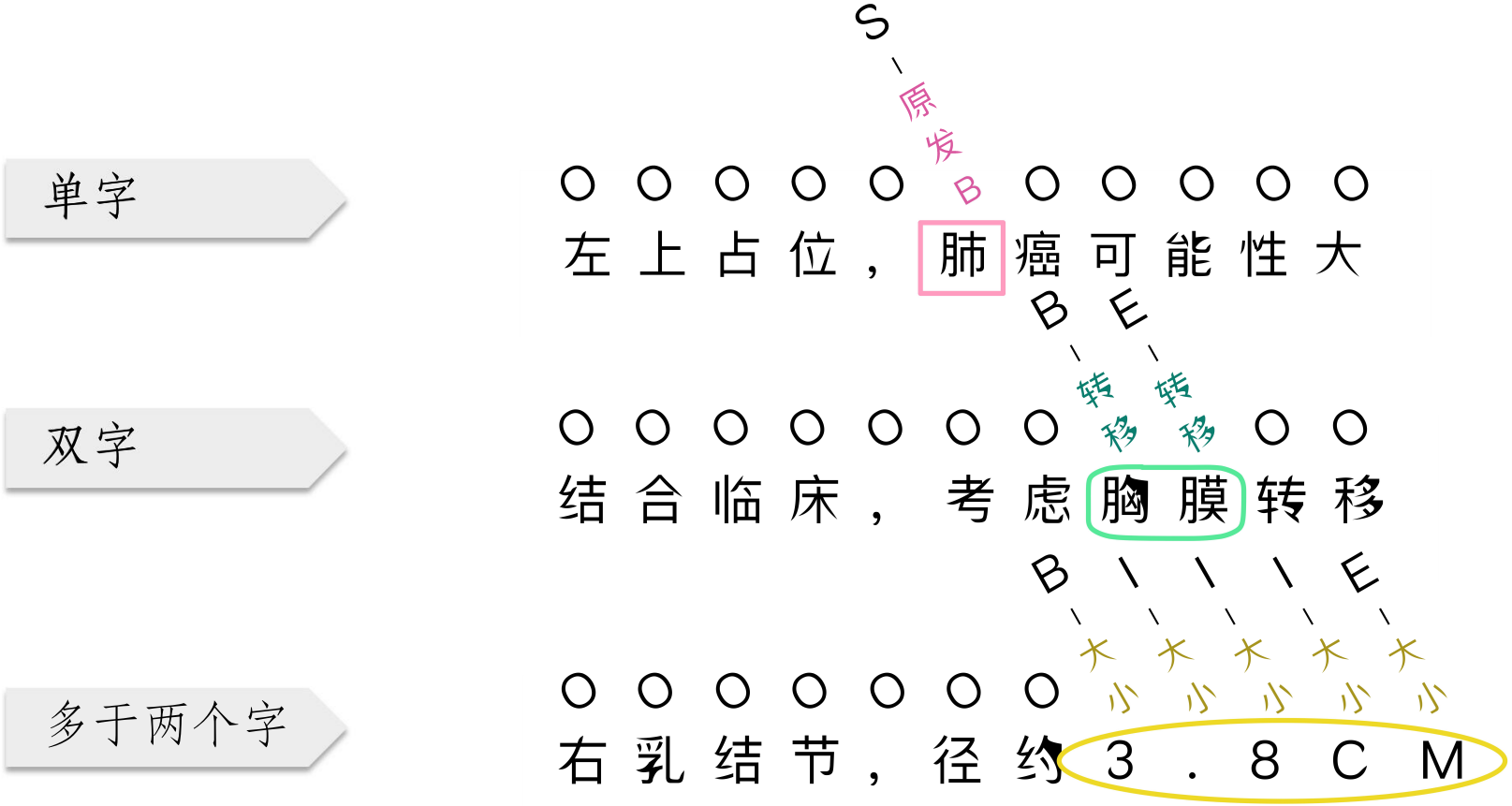


模型输出：BIEOS标签体系

- 4个属性槽位：肿瘤原发部位A、肿瘤原发部位B、原发部位大小、转移部位
- 5种位置标签：BIEOS(Begin，Inside，End，Outside，Single)

B-肿瘤原发部位A	**I**-肿瘤原发部位A	**E**-肿瘤原发部位A	**S**-肿瘤原发部位A	O
B-肿瘤原发部位B	**I**-肿瘤原发部位B	**E**-肿瘤原发部位B	**S**-肿瘤原发部位B	
B-原发病灶大小	**I**-原发病灶大小	**E**-原发病灶大小	**S**-原发病灶大小	
B-转移部位	**I**-转移部位	**E**-转移部位	**S**-转移部位	

模型输出：BIEOS标签体系

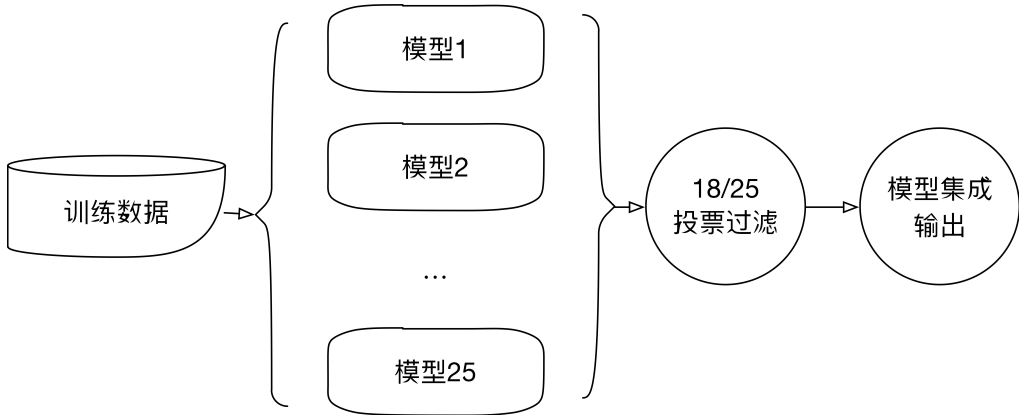


其他策略

- EMA(指数滑动平均)

$$EMA_t = \begin{cases} x_0 & t = 0 \\ \alpha x_t + (1-\alpha) EMA_{t-1} & t > 0 \end{cases}$$

- 模型集成



策略效果总览

模型	平均F1	消融影响
RoBERTa	76.79	N/A
参评系统	79.47	N/A
- CRF	77.98	1.49
- EMA	78.21	1.26
- 领域适配&任务适配	78.43	1.04
- 任务适配	79.01	0.46
- 领域适配	79.09	0.38
- 回译数据增强	79.24	0.23
- 关键词特征	79.28	0.19

- 训练集上5折交叉验证
- 单模型
- 未进行后处理

目 录

CONTENT

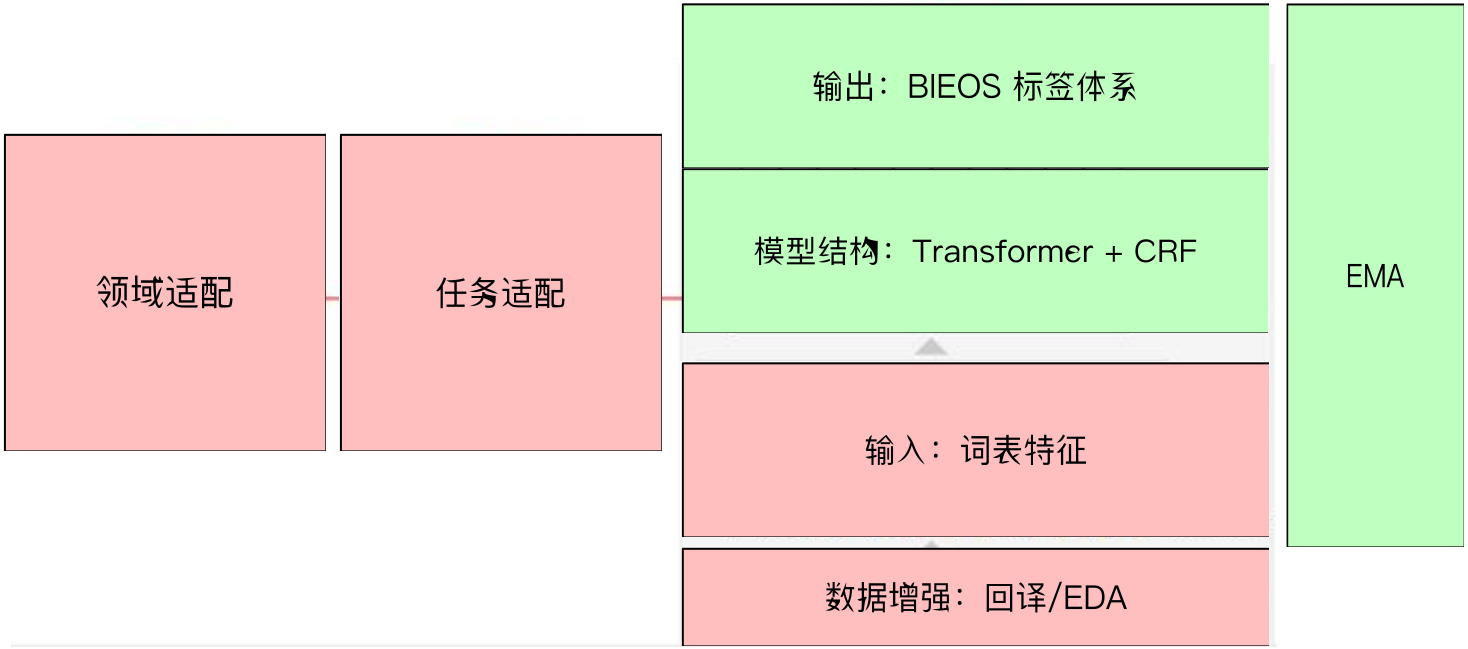
后处理策略

- **过滤**掉没有相应肿瘤原发部位的原发病灶大小
- 为了避免将器官大小误认为病灶大小，**过滤**掉没有关键提示文字能够确认此尺寸属于“病灶”、“影像密度影”或“B 超回声区”的原发病灶大小
- **过滤**掉没有关键提示文字“转移”的转移部位
- 若预测出多个肿瘤原发部位，仅输出多个模型中出现频率最高的作为最终答案
- 若「肿瘤原发部位A」与「肿瘤原发部位B」有文本上的重合，过滤掉「肿瘤原发部位B」

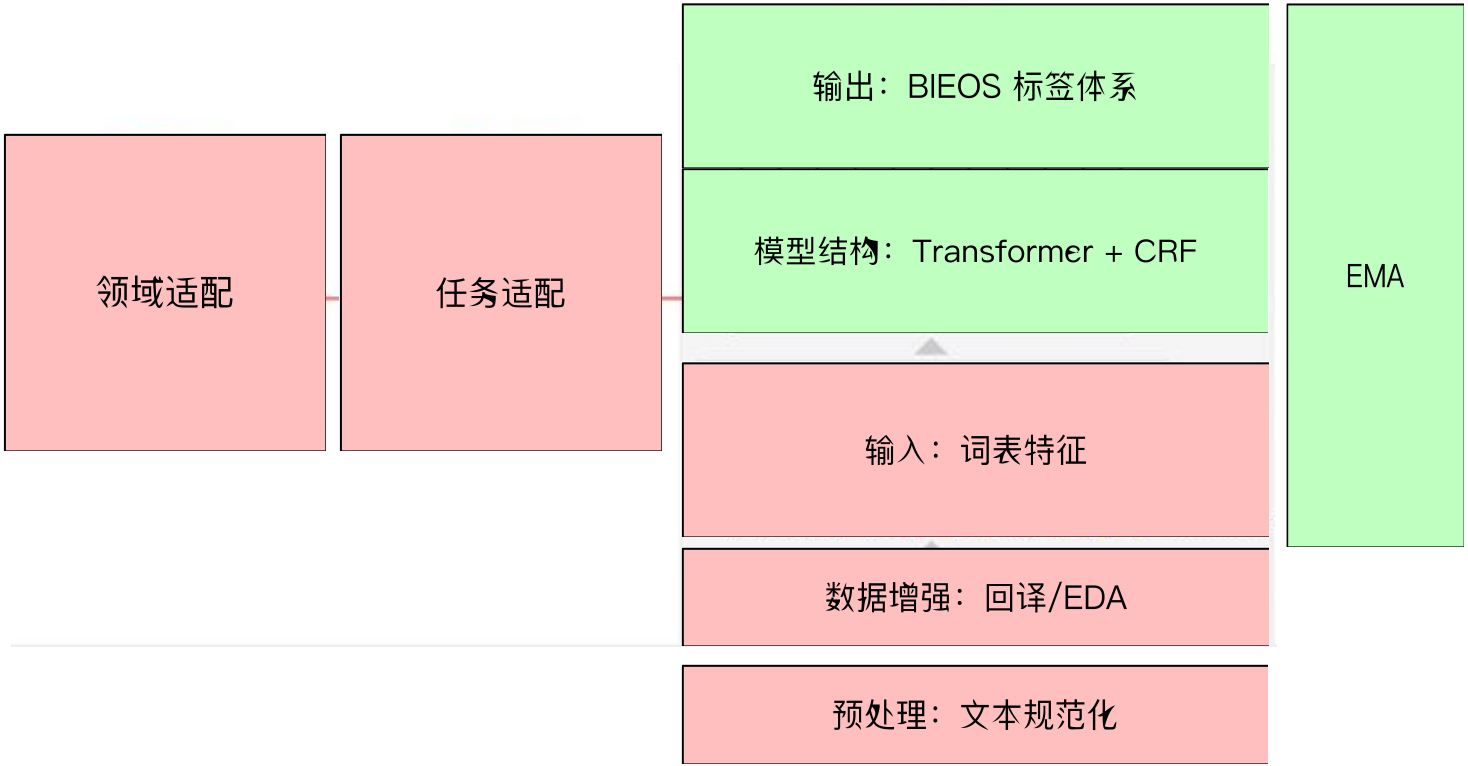
目 录

CONTENT

小样本下的医疗事件抽取系统



小样本下的医疗事件抽取系统



THANKS