基于预训练语言模型的小样本医疗事件抽取

戴松泰, 王泉, 黄苹苹, 吕雅娟, 朱勇

百度 知识图谱部

{daisongtai, wangquan05, huangpingping,
lvyajuan, zhuyong}@baidu.com

摘要. 医疗事件抽取是指从无结构医疗文本中抽取指定事件属性。它是临床电子病历结构化的重要步骤，具有较高的研究价值。2020 年全国知识图谱与语义计算大会 (CCKS 2020) 设置了医疗事件抽取任务评测，要求参评系统从临床电子病历中抽取肿瘤事件的三种指定属性，包括肿瘤原发部位、原发病灶大小、转移部位。本文介绍百度知识图谱部的参评系统。该系统借助预训练语言模型，通过领域适配、任务适配、任务精调实现小样本条件下的医疗事件抽取，以测试集 F1 76.23% 的最终成绩，在评测中排名第一。

关键词: 属性抽取 · 电子病历 · 预训练语言模型.

1 引言

近年来，随着自然语言理解技术的飞速发展以及医疗信息化建设的广泛应用，如何利用自然语言理解技术对海量医疗文本进行处理成为了一个越来越受到关注的问题。医疗事件抽取旨在从无结构医疗文本中抽取指定事件属性，是医疗文本结构化的重要步骤之一，也是临床上病历质控、辅助诊断等广泛应用的基础。当前医疗事件抽取任务普遍面临标注成本过高、标注样例稀少的问题。因此，在小样本条件下实现高质量的医疗事件抽取，对医疗事件抽取技术的广泛应用具有重要价值。2020 年全国知识图谱与语义计算大会 (CCKS 2020) 设置了医疗事件抽取任务评测，要求参评系统利用有限标注数据，从临床电子病历中抽取肿瘤事件的三种指定属性，包括肿瘤原发部位、原发病灶大小、转移部位。

基于 RNN[11] 及其变体 (如 LSTM[5]、GRU[1] 等) 的方法在医疗事件抽取相关任务上取得了较好的效果。而近年来以 BERT[3] 为代表的预训练语言模型，在多项自然语言理解任务上都取得了很大进步 [3, 16, 9, 13]，并展现了较强的泛化能力和领域迁移能力 [8]。参评系统借助预训练语言模型，

并综合利用多种策略，来提高小样本条件下的医疗事件抽取能力。具体地，参评系统在预训练流程里加入领域适配和任务适配 [4]，提升了语言模型对任务文本的建模能力，让语言模型在精调阶段的小样本下有更好的表现。其次，参评系统利用回译 (back translation) [12] 的方法进行数据增强，弥补了训练数据有限的问题。最后，参评系统还将实体词表作为关键特征加入到模型输入中，提升了模型在小样本条件下下对答案的拟合能力。最终参评系统以测试集 F1 76.23% 的最终成绩，在评测中排名第一。

2 问题

2.1 问题定义

表 1. 肿瘤事件属性抽取数据示例 [17]

原文：右肺癌化疗后，对比 2016-11-29CT：右上肺病变较前范围稍缩小，周边少许炎症较前稍减少。两肺散在小结节，大致同前。**左侧锁骨下区、纵隔多发淋巴结**，考虑转移，较前稍缩小。肝囊肿。左肾小囊肿。右肺癌化疗后，对比 2016-11-29CT：**右肺上叶**见不规则结节状、片状病灶，边界不清，最大层面大小约 **12mm×8mm**，边缘呈分叶状，增强扫描不均匀强化，紧贴斜裂胸膜，部分范围较前略缩小，右上肺见少许斑片状稍高密度影，边界不清，较前明显减少。左下肺（se8，im96）、左上肺（se8，im221）及右下肺（se8，im104）散在数个小类结节，边界清，大者直径 3mm，大致同前。右肺上叶前内基底段支气管变窄，基底段支气管分支管壁增厚，气管及其余支气管分支通畅。
肿瘤原发部位：右肺上叶 原发病灶大小：12mm×8mm 转移部位：左侧锁骨下区、纵隔多发淋巴结

如表1所示，对于给定主实体为肿瘤的电子病历文本数据，任务目标是抽取出预定义的三种肿瘤事件属性（肿瘤原发部位，原发病灶大小和转移部位）。其中，肿瘤原发部位指肿瘤最先发生于的组织或者器官，原发病灶大小指原发部位的大小，转移部位指肿瘤从最先发生的组织或者器官转移到的其他组织或器官。

2.2 评价指标

任务使用基于属性实体的微平均 F1 值作为评测指标，即评测指标使用属性实体作为基本单位来计算相应准确值、召回值和 F1 值。

3 方法

3.1 数据预处理

参评系统首先采用数据清洗、长文本切分、答案回标等策略对主办方提供的评测数据进行预处理，详细描述如下。

数据清洗 该步骤旨在去除官方评测数据中的非法字符，并使其规范化。具体的清洗策略包括：

1. 将全角数字和字母转换成半角，英文字母统一转换成大写。
2. 去除特殊字符，如 ^@，^A，^Z, ^? 等，去除空格。
3. 将癌症相关的英文缩写替换成中文，如 cancer–> 癌，CA–> 癌，MT–> 癌，carcinoma–> 癌等。[1]

长文本切分 该步骤在保证句子完整的前提下对超长的输入文本进行切分，每个切分后的文本片段长度不超过 510 个字符，从而满足绝大部分预训练语言模型对于输入文本的长度限制。[2]

答案位置回标 官方训练数据针对目标槽位直接提供相应答案。参评系统采用序列标注模型来建模槽位填充任务，需要标注答案在输入文本中的位置。一般通过精确匹配直接进行答案位置回标。对于少量匹配不上或匹配到多个候选的情况，进行如下特殊处理：

– 对于*原发病灶大小*这一槽位，通过正则表达式适当放宽匹配限制，例如允许正确答案「5.1 CM × 2.7 CM」匹配输入文本中的「5.1 * 2.7 CM」
– 对于*转移部位*这一槽位，若匹配到输入文本中的多个候选，选择离*转移*两字最近的候选作为最终答案并标注其位置
– 对于*肿瘤原发部位*这一槽位，若匹配到输入文本中的多个候选，将与*原发病灶大小*位于同一句子中的候选作为最终答案并标注其位置

此外，通过观察数据发现*肿瘤原发部位*这一槽位表述方式多样，为其识别带来了较大困难。因此，参评系统特别引入*肿瘤位置*这一辅助槽位来帮助*肿瘤原发部位*的识别。辅助槽位*肿瘤位置*的标注方式如下：

[1] 因为观察到测试集中频繁出现的癌症相关英文缩写在训练集中大多直接以「癌」的形式出现，该转化可以帮助弥合训练测试不一致。

[2] 加上起始 [CLS] 和结尾 [SEP]2 个特殊字符，总长度不超过于 512。

– 对于那些匹配肿瘤原发部位但与原发病灶大小不在同一句中的提及，如果它与“癌”、“恶性”等关键词位于同一句子，则将其标记为肿瘤位置这一槽位的答案
– 对于那些匹配“X 癌”的提及，如果它与肿瘤原发部位没有文本上的重合（如“肾”vs.“肝右叶”），则将其标记为肿瘤位置这一槽位的答案[3]

3.2 模型训练

参评系统以中文预训练语言模型 RoBERTa-wwm-ext-large[2] 为基础，搭建序列标注模型以实现槽位填充。图1展示了参评系统的整体框架，包括领域适配、任务适配、任务精调三个阶段，分别介绍如下。

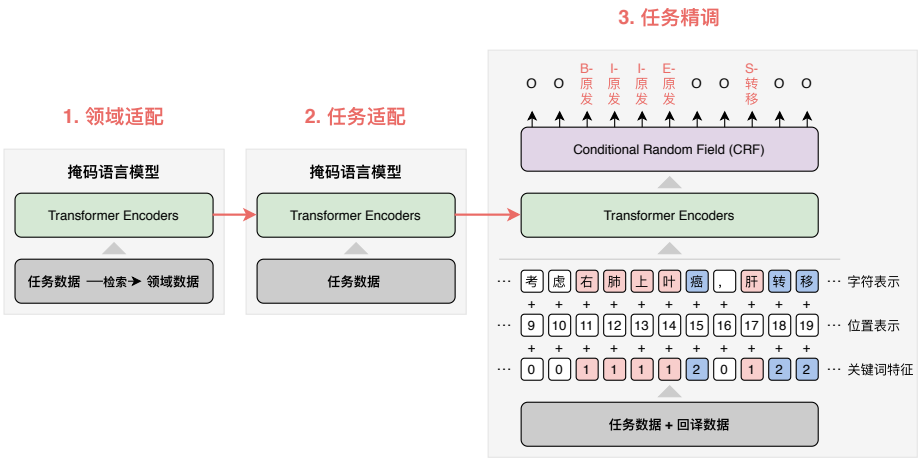

图 1. 参评系统的整体框架

领域适配 领域适配 [4] 是指在癌症与影像学领域的文本上进行语言模型预训练，旨在让语言模型更好地适应领域数据。具体地，参评系统采用如下策略获取领域数据：

1. 将官方提供的训练文本切分成 5-10 字符的短语，使用百度和必应搜索引擎在医学专业网站（如“好大夫在线”、“丁香园”、“影像 PPT”等）搜索上述短语，保存搜索引擎返回的前 50 条网页结果

[3] 肿瘤位置可以作为肿瘤原发部位的补充，在预测中起到查缺补漏的作用

2. 爬取并清洗上述搜索结果网址的网页全文, 并按句切分成长度不超过 510 字符的段落

最终获得约 96 万医学领域的文本段落，总计包含约 2.6 亿字符。在该领域数据上进行总计 10 轮掩码语言模型（masked language model）[3] 预训练。

任务适配 任务适配 [4] 是指领域适配之后，在官方提供的训练文本（包含标注文本和未标注文本, ）上进行语言模型预训练。我们将主办方提供的所有电子病历文本（包括标注的与未标注的）处理成小于 510 长度的文本片段（超长的文本按句号切分），共 4710 条。在此数据上进行总计 100 轮掩码语言模型预训练，进一步提升语言模型对任务数据的适应性和建模能力。

任务精调 参评系统采用 RoBERTa+CRF 的模型结构实现槽位填充，在 RoBERTa 模型 [9]24 层 Transformer 编码器 [14] 的基础上叠加条件随机场（conditional random field, CRF）[7] 进行序列标注。Transformer 编码器参数由领域适配和任务适配后的 RoBERTa 模型载入，其余参数在任务训练数据上精调得到。

模型输入 对于输入文本中的任意字符,首先利用字符表征(token embedding)和位置表征（position embedding）对其进行表示。此外，引入**关键词特征（keyword feature）**标示那些对答案有较强提示作用的关键词。这里关键词包括两种:

1. 官方提供的实体词表以及利用简单规则对词表进行的扩充，例如将“左肺中叶”扩充为“左肺上叶”, “左肺下叶”, “右肺上叶”, “右肺中叶”, “右肺下叶”等
2. 人工补充的个别关键词，如“癌”, “转移”, “恶性”等

图1给出了关键词特征示例，其中 0 表示非关键词，1 表示官方实体词表（及其扩充），2 表示人工补充关键词。

标签体系 参评系统采用 BIEOS（Begin, Inside, End, Outside, Single）标签体系 [15] 进行序列标注。总共涉及四种类型的槽位，包括*肿瘤原发部位*、*原发部位大小*和*转移部位*三种官方指定槽位，以及*肿瘤位置*这一辅助槽位。每种槽位设有单独的 BIES 标签（如 B-肿瘤原发部位、I-肿瘤原发部位、E-肿瘤原发部位），槽位之间共享 O 标签，总计 17 种标签。图 1 也给出了该标签体系示例。

此外, 参评系统采用了**回译（back translation）策略** [12] 进行数据增强, 利用百度翻译开放 API[4]将训练数据翻译成英文再回译成中文，从而将训练数据扩充 1 倍。参评系统在训练过程中还采用了**指数滑动平均（exponential moving average, EMA）** [10] 来提升模型的效果和稳定性。最后，参评系统采用了**模型集成（ensemble）策略**，在同样配置下，用不同的随机种子训练得到 25 个模型，以 18/25 的阈值，对答案进行投票过滤，以进一步提升系统效果。

3.3 后处理

最后，参评系统采用如下后处理策略，对模型预测答案进行过滤，输出最终答案:

- 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
- 若辅助槽位肿瘤位置与肿瘤原发部位没有文本上的重合，补充肿瘤位置为额外的肿瘤原发部位

4 实验

4.1 数据

CCKS2020 医疗事件抽取评测任务提供了 1000 条标注训练数据、400 条标注验证数据、300 条标注测试数据、1300 条未标注数据和 863 个实体词表。我们将主办方提供的 1000 条标注的训练数据划分成 5 份, 在 5 折（fold）交叉验证下实验并选择超参数。

4.2 实验设置

参评系统在 PaddlePaddle[5]框架上实现，PaddlePaddle 是百度开发的端到端开源深度学习框架。

[4] https://fanyi-api.baidu.com/product/12
[5] https://github.com/PaddlePaddle/Paddle

表 2. 实验超参数设置

超参数	领域适配	任务适配	精调 (参评系统)	精调 (RoBERTa)
批大小（Batch Size）	48	48	16	16
训练轮数（Epochs）	10	100	32	32
最大学习率（Peak Learning Rate）	3e-5	3e-5	6e-5	6e-5
学习率下降方法	线性	线性	线性	线性
学习率预热比例（Warmup Ratio）	0.05	0.05	0.1	0.1
最大序列长度（Max Sequence Length）	512	512	512	512
权重衰减（Weight Decay）	0.01	0.01	0.01	0.01
CRF 层学习率	N/A	N/A	1e-3	N/A

领域适配阶段 我们在 48 的批大小（batch size）下训练 10 轮（epoch）。训练中我们使用学习率预热（Warm up）与线性衰减（linear Decay）策略：前 5% 的 step 中学习率从 0 线性增长至 3e-05，后 95% 的 step 中学习率从 3e-5 线性衰减至 0。最大序列长度为 512，优化算法为 Adam 算法，权重衰减（weight decay）参数为 0.01。

任务适配阶段 我们采用与领域适配相同的参数，在任务文本上训练 100 轮（epoch）。

精调阶段 在 1000 条标注的训练数据的 5 折交叉验证下，我们在 {1.5e-5, 3e-5, 6e-5, 1e-4} 的范围内调学习率，在 {4, 8, 16, 32} 的范围内调整批大小 (batch size)，在 {4, 8, 16, 32, 64, 128} 的范围内调试训练轮数 (epoch)，最优参数如表2所示。

4.3 实验结果

我们以中文 RoBERTa-wwm-ext-large[2] 作为基线与参评系统进行对比。各模型在训练集上进行 5 折交叉验证 [6]，每份重复运行 5 次，得到 25 次实验的平均 F1 值（公式1）。对每一份（fold）数据下 5 次重复实验结果的标准差求平均，得到 5 折交叉验证的平均标准差（公式2）。

$$F1_{mean} = \frac{\sum_{i=0}^{4}\sum_{j=0}^{4} F1_{i,j}}{25} \tag{1}$$

$$SD_{mean} = \frac{\sum_{i=0}^{4} SD_i}{5} \tag{2}$$

表 3. 模型效果对比

模型	平均 F1	平均标准差
RoBERTa[2]	76.79	0.61
参评系统	79.47	0.46
- CRF	77.98	0.80
- 领域适配	79.09	0.68
- 任务适配	79.01	0.58
- 领域适配 & 任务适配	78.43	0.77
- 关键词特征	79.28	0.78
- 回译数据增强	79.24	0.71
- EMA	78.21	0.68

从表3中可以看出，条件随机场（Conditional Random Field, CRF）和指数滑动平均（expenetial moving average, EMA）对模型的增益最大。而语言模型的领域适配与任务适配对任务都有独立的增益，说明领域适配和任务适配从不同侧面提升了模型对任务文本的拟合能力。此外，关键词特征的引入和基于回译的数据增强对模型效果也有一定的提升作用。

5 结论

本文介绍了一种医疗事件抽取系统，相比于直接使用预训练语言模型，该系统通过领域适配、任务适配、回译数据增强以及加入关键词信息的方式在小样本条件下的医疗事件抽取中获得了更好的效果。在 2020 年全国知识图谱与语义计算大会（CCKS）评测中取得了 76.23% 的 F1 分数，在排行榜位列第一。在评测中，我们发现在验证集上有效的一些方法和特征，在测试集上效果并不显著，经过数据分析，我们发现这是由于测试集文本分布与验证集和训练集分布不同导致的。由于训练集较小，模型容易学到一些泛化能力较差的特征。在未来的工作中，我们的会进一步提升模型在有限训练集上学习的泛化能力和迁移能力。

参考文献

[1] J. Chung et al. “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”. In: *ArXiv* abs/1412.3555 (2014).

[2] Yiming Cui et al. “Revisiting Pre-Trained Models for Chinese Natural Language Processing”. In: *Findings of EMNLP*. Association for Computational Linguistics, 2020.

[3] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT*. 2019.

[4] Suchin Gururangan et al. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *Proceedings of ACL*. 2020.

[5] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[6] Ron Kohavi. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. In: *IJCAI*. 1995.

[7] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *ICML*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. ISBN: 1558607781.

[8] Nelson F. Liu et al. “Linguistic Knowledge and Transferability of Contextual Representations”. In: *NAACL-HLT*. 2019.

[9] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[10] Chao-Wen Lu and Marion R. Reynolds Jr. “EWMA Control Charts for Monitoring the Mean of Autocorrelated Processes”. In: *Journal of Quality Technology* 31.2 (1999), pp. 166–188. URL: https://doi.org/10.1080/00224065.1999.11979913.

[11] D. Rumelhart, Geoffrey E. Hinton, and R. J. Williams. “Learning internal representations by error propagation”. In: 1986.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Improving Neural Machine Translation Models with Monolingual Data”. In: *ACL*. 2016. URL: https://www.aclweb.org/anthology/P16-1009.

[13] Y. Sun et al. “ERNIE 2.0: A Continual Pre-training Framework for Language Understanding”. In: *AAAI*. 2020.

[14] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. 2017.

[15] Y. Xia and Q. Wang. “Clinical Named Entity Recognition : ECUST in the CCKS-2017 Shared Task 2”. In: 2017.

[16] Zhilin Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Advances in Neural Information Processing Systems 32*. 2019, pp. 5753–5763.

[17] 面向中文电子病历的医疗实体及事件抽取技术评测任务书. URL: http://sigkg.cn/ccks2020/wp-content/uploads/2020/03/3-CCKS2020技术评测-面向中文电子病历的医疗实体及事件抽取.docx.