基于 BERT 的电子病历实体关系联合抽取研究*

黄晓芳 陈剑秋 周祖宏 廖 敏

（西南科技大学计算机科学与技术学院 绵阳 621000） （绵阳市中心医院 绵阳 621000）

〔摘要〕 分析中文电子病历数据实体关系提取常用方法，提出一种基于双向编码器表征的实体关系联合抽取算法，使用级联解码器以及指针标注方法完成实体关系抽取及实体识别，实验结果证明该方法可有效抽取电子病历实体关系。

〔关键词〕 电子病历；关系抽取；联合抽取模型；自然语言处理

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.05.005

Study on Joint Extraction of Chinese Electronic Medical Record Entity Relationship Based on BERT *HUANG Xiaofang, CHEN Jianqiu, School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621000, China; ZHOU Zuhong, LIAO Min, Mianyang Central Hospital, Mianyang 621000, China*

〔**Abstract**〕 The paper analyzes the common methods of entity relationship extraction for Chinese electronic medical record (EMR) data, proposes a joint extraction algorithm of entity relationship based on bidirectional encoder representation from transformers (BERT), which uses cascade decoder and pointer annotation method to complete entity relationship extraction and entity recognition. Experimental results show that the algorithm used in the paper can effectively extract the entity relationship of electronic medical record.

〔**Keywords**〕 electronic medical record (EMR); relationship extraction; joint extraction model; natural language processing (NLP)

1 引言

电子病历（electronic medical records，EMR）是医务人员使用电子医疗系统产生文字、符号、图表、图形、数据和影像等数字化信息，并将其进行存储所形成的医疗记录[1]。EMR 不仅包括患者临床信息，如检查结果、临床诊断以及不良反应等，还包括丰富的医疗实体[2]。实体关系抽取是自然语言处理（natural language processing，NLP）信息抽取技术中的基本任务，也是构建知识库和知识图谱的关键方法[3]。从 EMR 文本中挖掘医疗实体以及实体间的语义关系，对于推动 EMR 在医疗健康服务中的应用具有重要意义。

电子病历文本作为医学文本的重要组成部分，是构建医学知识图谱的基础。相比通用领域，电子病历文本的特点主要为实体的高密度分布以及实体间关系的交叉互联，这一特点可能造成实体嵌套与实体重叠等问题。如“主动脉夹层”为一种疾病实

〔修回日期〕 2022－11－08

〔作者简介〕 黄晓芳，博士，教授。

〔基金项目〕 四川省科技厅重点研发项目“面向多语种的全球前沿技术智能汇集与应用平台”（项目编号：2021YFG0031）。

体，而该疾病实体中包含“主动脉”这一部位实体，这类实体就被称为重叠实体。例如，句子中有“同义词”与“治疗手段”两个关系类型，“局限性蔓状血管瘤”是关系“同义词”的头实体，也是关系“治疗手段”的头实体，这类三元组被称为实体重叠三元组。准确处理嵌套实体与实体重叠三元组是电子病历文本信息抽取的难点，见图 1。

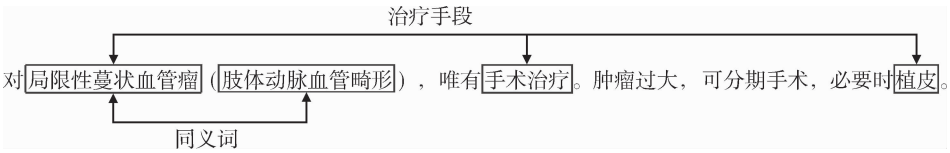

图 1 实体重叠三元组举例

本文以中文电子病历数据作为研究对象，提出一种基于双向编码器表征（bidirectional encoder representations from transformers，BERT）的实体关系联合抽取方法，使用级联解码器以及指针标注方法完成实体关系抽取及实体识别，有效解决联合抽取模型的实体冗余以及在抽取实体重叠三元组时带来的错误传播等问题。

2 相关工作

2.1 常用关系抽取方法

2.1.1 *单任务学习方法* 目前在生物医学关系抽取方法中主要采取单任务学习方法即流水线学习[4]，学习过程相对独立。在采取该方法进行模型提取时，命名实体识别与实体关系抽取两个任务分开进行，虽然方便调参，子任务之间耦合度低，但考虑两个问题之间的内在联系，会出现错误传播、交互缺失等问题。

2.1.2 *多任务学习方法* 即联合学习，就是利用任务间的相关信息来提升模型性能。在合理的可用性约束下，尽可能提高模型的识别精度。联合抽取模型将两个子模型统一建模，可以进一步利用两个任务之间的潜在信息[5]。联合学习主要包括共享参数[6]和联合解码[7]两类。共享参数联合学习模型通过共享参数层来进行实体关系抽取。例如 Miwa M[8]等提出将两个子任务联合到一个模型中，直接对三元组进行建模，先进行实体识别再进行关系抽取，通过实体及关系共同更新网络参数。但该方法会抽取大量未能组成三元组的实体，造成实体冗余。Yu B[9]等提出一种新颖的分解策略，首先抽取头实体，再抽取其每个关系相应的尾实体，该方法只识别可能形成三元组的头实体，从而减轻冗余实体的影响，但该方法对嵌套实体以及实体重叠三元组的学习能力较差。联合解码通过共享解码层实现联合，加强两个子任务之间的交互。

2.2 方法选取

2.2.1 *已有实体重叠三元组关系抽取方法* 针对实体重叠三元组关系抽取任务，Zeng X[10]等提出一种基于 Copy 机制的端到端模型，根据实体关系创建多个解码器，从原句中复制出实体。Fu T J[11]等提出一种端到端的模型来进行实体关系联合抽取，利用图形卷积网络来共同学习命名的实体和关系，然后基于图的方法，对实体重叠三元组进行预测。但是以上两种方法都是将关系类别当作离散的标签来进行分类，针对关系分类器进行优化，没有从分类器层面对实体重叠三元组进行改进学习。

2.2.2 *本研究选用的抽取方法* 本文采用 Yu B[9]等提出的先抽取头实体、再针对每个关系提取其相应的尾实体的实体关系抽取策略，但电子病历文本存在实体高密度分布以及实体间关系交叉互联的特点，单纯采用这种实体抽取策略并不能很好地解决问题。针对电子病历文本的特点，首先使用指针标注的方法进行实体识别，解决电子病历文本中存在嵌套实体的问题，在此基础上改进尾实体关系特定标注器来清除实体重叠三元组对模型的影响，提高模型在处理电子病历文本时的性能。

3 研究方法与过程

3.1 基于 BERT 的实体关系联合抽取模型

基于 BERT 的实体关系联合抽取模型采用参数共享的方法进行实体关系联合抽取，包括词嵌入模块、头实体抽取模块以及关系特定标注器模块，见图 1。对于中文电子病历文本，首先通过词嵌入模块将其转换为矩阵向量，其次通过头实体标注器将句子中可能存在的头实体全部标注出来，最后通过关系特定尾实体标注器，对于每个找到的头实体，使用每个关系的特定标注器找到相关关系和对应尾实体。

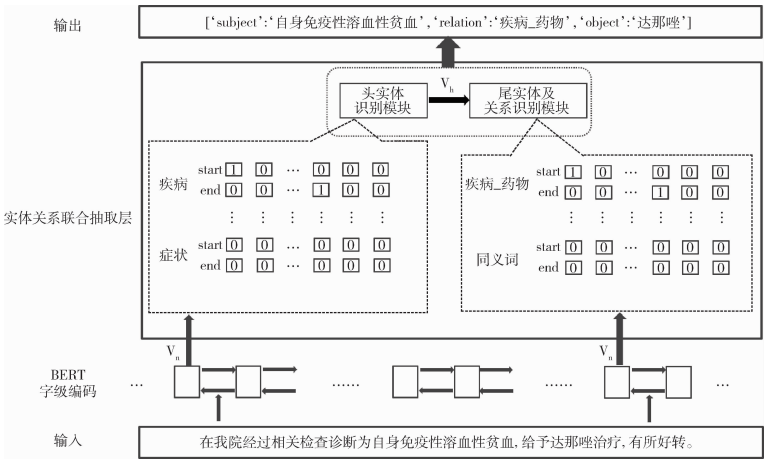

图 2 模型整体框架

3.2 词嵌入模块

本文采用参数共享的方式进行实体关系联合抽取，参数共享采用共享词向量的方式进行，其中输入字符模块中 CLS 用于后续的分类，SEP 用于分开两个输入句子。词嵌入模块将 BERT - WWM[12] 作为词嵌入器，见图 3。

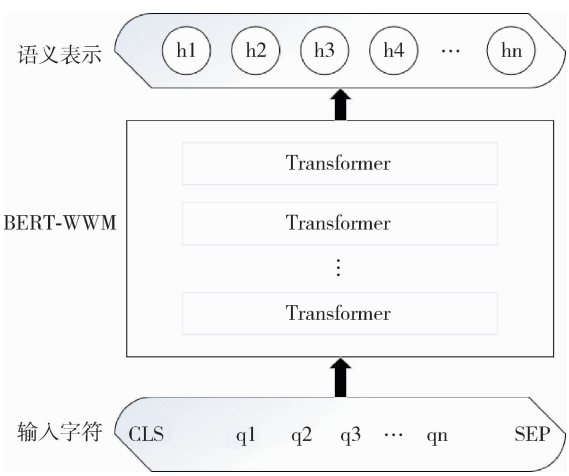

图 3 BERT - WWM 模块

采用与 BERT - base - Chinese 一样的模型架构，由 12 层 Transformer 构成，相较于 BERT - base - Chinese，BERT - WWM 在预训练阶段的掩码语言模型（mask language model，MLM）任务中，按照中文的语言习惯对词进行遮罩，提高了模型在处理中文自然语言处理任务时的性能。

3.3 实体关系联合抽取模块

3.3.1 *头实体识别模块* 头实体识别模块用来识别所有可能的头实体，针对中文电子病历文本存在的实体嵌套问题采用了指针标注的方法。相较于 Miwa M[8] 等提出的序列标注方法，见图 4，模型默认一个 token 只有一个标签，而该方法在处理嵌套实体时，会出现一个 token 对应多个标签的情况，实际上就将问题转换为了多标签分类[12]问题，而嵌套实体在医学实体中占比很小，又会造成嵌套实体数据与非嵌套实体数据长尾分布[13]的问题。针对序列标注的这一弱点，本文采用指针标注方法，见图 5，解决了序列标注在处理实体嵌套时出现的单标签多分类转换为多标签分类的问题。指针标注针对每个实体类别创建实体头尾标注器，弥补了序列标注方法在处理嵌套实体时存在的不足，B 代表头实

体的开始，I 代表头实体的中间片段以及结尾，头实体的开始和结束都用 1 来表示，非边界字符用 0 表示。

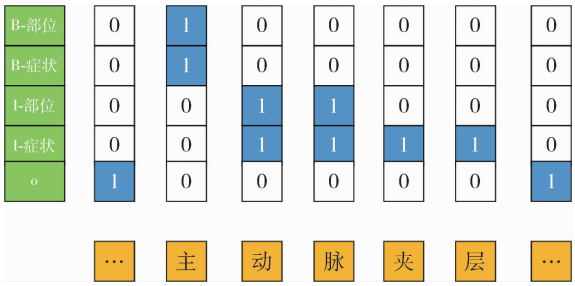

图 4 序列标注

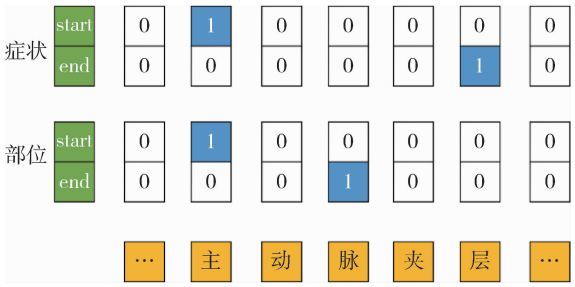

图 5 指针标注

在采用指针标注时，每一层都构建一个 sentence_ len * 2 的矩阵，也就是针对每个 token 都有 $N*2$ 个二分类（N 为实体类别个数），用来判断该索引位置的标签，每个类型头实体边界的概率计算方法如下：

$$p_i^{start\_s} = \sigma(W_{startX_i} + b_{start}) \quad (1)$$

$$p_i^{end\_s} = \sigma(W_{endX_i} + b_{end}) \quad (2)$$

P_i 表示输入的句子中第 i 个 token 是头实体的头或尾的概率。为使模型的泛化功能更佳，使用全连接层进行计算，其中 W 表示权重，b 表示偏置，σ 是输出层激活函数，X_i 表示文本中第 i 个字符的编码表示。将计算结果与阈值进行比较，超过阈值则置为 1，否则为 0。如果识别到多个头实体，那么就将识别到的 start 和 end 按就近原则组合成一个实体。

3.3.2 尾实体及关系识别模块 关系特定标注器模块采用 Yu B[9] 等提出的 ETL – BIES 模型进行改进，ETL – BIES 通过关系名称来对尾实体进行标注，采用指针标注的尾实体标注器（其中 MED 为关系药物治疗的标签），该标注方法在处理实体重叠三元组（如句子“患者初步诊断多发性肌炎，该疾病一般以对称性肌无力为特征，该患者发病初期特征不明显，但后期以并发症出现。”中可提取三元组：“‘多发性肌炎’—‘临床表现’—‘对称性肌无力’”与“‘多发性肌炎’—‘传播途径’—‘对称性肌无力’”）时效果不佳，也会造成将单标签多分类转换为多标签分类，从而影响模型性能，见图 6。

针对中文电子病历文本存在重叠实体问题，提出一种改进的关系特定尾实体标注器模块。该模块包含多个关系特定尾实体标注器，每个关系对应一个特定的尾实体标注器。模型能够充分捕捉到头实体与尾实体以及关系之间的语义关系，解决实体冗余的问题，关系特定的尾实体标注器也采用指针标注的方法，由尾实体的开始位置和结束位置组成，见图 7。针对每一个提取到的头实体，遍历所有的关系 r，判断头实体和特定的关系 r 是否存在尾实体，尾实体边界概率的计算方法如下：

$$p_i^{start\_o} = \sigma(W_{start}^r(X_i + V_{sub}^k) + b_{start}^r) \quad (4)$$

$$p_i^{end\_o} = \sigma(W_{end}^r(X_i + V_{sub}^k) + b_{end}^r) \quad (5)$$

图 6 尾实体标注器

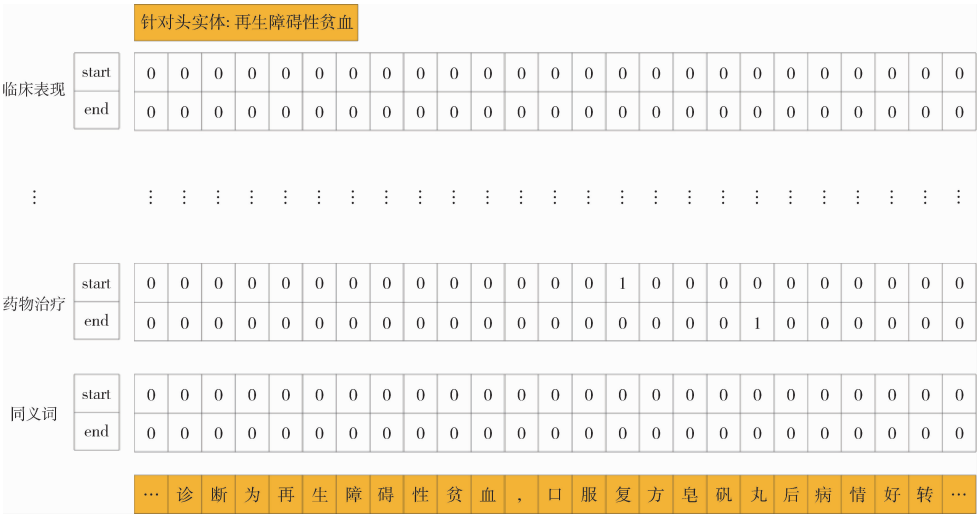

图 7 关系特定尾实体标注器

关系特定尾实体标注器采用全连接层进行计算，与头实体标注器不同的是在计算语义向量时要联合头实体的语义信息，其中 V_{sub}^{k} 表示头实体标注器识别到的第 k 个头实体编码，X_i 是句子第 i 个字符的编码表示，其中 W 表示权重，b 表示偏置，σ 是输出层激活函数。针对每个头实体，对于所有关系重复在整个句子 X 上使用上述公式计算，从而为每个头实体找出每个关系下可能存在的尾实体。该模型的损失函数均采用二分类交叉熵损失函数：

$$L = \frac{1}{N}\sum_{i} -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \quad (6)$$

4 实验与结果分析

4.1 数据集

采用第六届中国健康信息处理会议（China Health Information Processing Conference，CHIP）2020 测评任务 3 中基于 schema 的中文电子病历信息抽取数据集。该数据集共包含 44 个种类的实体关系，其中部分种类三元组数量较少，导致整个数据集存在长尾分布的问题。本文选取数量最多的 6 种实体关系进行统计，最后组成的数据集共包含电子病历实体关系三元组 43 219 对，见表 1。再从各类实体关系中选取 80% 用于训练，20% 用于测试。

表 1 医疗实体关系类别分布

类别	三元组数量（对）
疾病_ 症状	14 113
疾病_ 疾病	9 456
疾病_ 社会学	5 540
疾病_ 药物	5 484
疾病_ 检查	5 319
疾病_ 治疗手段	3 307
合计	43 219

4.2 评估标准

选用关系抽取领域常用的评价指标：准确率（precision，P）、召回率（recall，R）和 F 值（F - score）来评估关系抽取的效果：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$\text{F} - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

其中，TP 表示测试集中该类别的正例被正确分类的统计值。FN 表示测试集中该类别的负例被错误分类的统计值。

4.3 实验结果

4.3.1 结果对比 除使用本文提出的模型外，还使用了 ETL - BIES[9]、ETL - BIES - WWM

(ETL – BIES 模型的词嵌入层更换为 BERT – WWM) 以及 BERT – MultiHead 做对比实验，见表 2、表 3。

4.3.2 采用不同词嵌入器的 ETL – BIES 实验结果分析 使用 BERT – WWM 作为词嵌入器的模型比使用 BERT – base – Chinese 作为词嵌入器在各个关系类别上表现更佳，因为 BERT – base – Chinese 在预训练阶段只能对中文文本进行字级别的遮罩，而 BERT – WWM 可以将中文文本进行词级别的遮罩，在中文语义环境下提供完整的语义信息，增强模型在训练过程中对中文语义信息的理解。

表 2 采用不同词嵌入器的 ETL – BIES 实验结果

关系	ETL – BIES			ETL – BIES – WWM		
	P	*R*	*F*	*P*	*R*	*F*
疾病_ 症状	0.696	0.640	0.655	0.673	0.642	0.658
疾病_ 疾病	0.618	0.628	0.622	0.622	0.634	0.628
疾病_ 社会学	0.578	0.433	0.505	0.582	0.434	0.508
疾病_ 药物	0.653	0.645	0.674	0.656	0.648	0.676
疾病_ 检查	0.650	0.641	0.645	0.656	0.643	0.650
疾病_ 治疗手段	0.630	0.618	0.624	0.638	0.621	0.630
整体关系类别	0.665	0.631	0.648	0.670	0.634	0.652

表 3 关系抽取结果

关系	ETL – BIES – WWM			BERT – MultiHead			本文提出的模型		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
疾病_ 症状	0.673	0.642	0.658	0.749	0.736	0.743	0.742	0.751	0.747
疾病_ 疾病	0.622	0.634	0.628	0.692	0.679	0.686	0.722	0.714	0.718
疾病_ 社会学	0.582	0.434	0.508	0.652	0.638	0.645	0.634	0.612	0.623
疾病_ 药物	0.656	0.648	0.676	0.688	0.679	0.684	0.692	0.688	0.690
疾病_ 检查	0.656	0.643	0.650	0.663	0.678	0.671	0.682	0.689	0.686
疾病_ 治疗手段	0.638	0.621	0.630	0.664	0.653	0.658	0.642	0.661	0.652
整体关系类别	0.670	0.634	0.652	0.686	0.684	0.685	0.708	0.686	0.696

4.3.3 关系抽取结果分析 3 个模型的性能在关系抽取上差别不大，整体关系类别上的 *F* 值都在 0.65 以上。因为本文提出的模型基于 ETL – BIES 针对实体嵌套与实体重叠分别做了改进，所以在不同类别上本文模型的 *F* 值均高于 ETL – BIES。从具体类别来看，语料越丰富的关系模型学习到的语义特征越优秀，在疾病_社会学与疾病_治疗手段上，BERT – MultiHead 的表现优于本文提出的算法，这其中可能有以下原因：一是疾病_社会学实体关系中存在病史、风险评估因素等子关系类别，在此类句子中实体间关系较远，而 BERT – MultiHead 中的多头选择机制[13]对于长文本的特征提取能力较强；二是疾病_治疗手段关系数量相对较少，模型对其特征的学习能力不足。除去疾病_社会学与疾病_治疗手段外，本文模型的 *F* 值均高于 ETL – BIES 与 BERT – MultiHead，这表明基于 BERT 的电子病历实体关系联合抽取模型在处理电子病历文本中的实体嵌套与实体重叠问题上是有效的。BERT – MultiHead 虽然能够获得较优的全局文本特征，但需要将 BERT 预训练模型的输出与多头选择层结合起来，两个模型无法充分拟合，因此在处理实体重叠等问题上效果相对较差。

5 结语

本文针对中文电子病历存在的实体嵌套以及实体重叠问题，采用参数共享的思想构造头实体标注器以及关系特定尾实体标注器，最后将标注结果整合成电子病历三元组进行输出，在 CHIP 2020 关系抽取任务中取得良好效果。该方法仍有一定的改进空间：一方面在实体类别以及实体关系类别较多时，会生成较多二分类器，可能会造成资源浪费等问题，影响模型效率；另一方面可以拓展语料库，如采用在大规模中文电子病历文本上预训练的

BERT 模型来获取专业领域特征。

参考文献

1 马锡坤，杨国斌，于京杰．国内电子病历发展与应用现状分析［J］．计算机应用与软件，2015，32（1）：4.

2 杨锦锋，关毅，何彬，等．中文电子病历命名实体和实体关系语料库构建［J］．软件学报，2016，27（11）：2725－2746.

3 HEIKO P. Knowledge graph refinement：a survey of approaches and evaluation methods［J］. Semantic web，2016，8（3）：489－508.

4 XU Y，MOU L，GE L，et al. Classifying relations via long short term memory networks along shortest dependency paths［C］. Lisbon：The 2015 Conference on Empirical Methods in Natural Language Processing（EMNLP），2015.

5 KATIYAR A，CARDIE C. Investigating LSTMs for joint extraction of opinion entities and relations［C］. Berlin：The 54th Annual Meeting of the Association for Computational Linguistics（ACL），2016.

6 KATIYAR A，CARDIE C. Going out on a limb：joint extraction of entity mentions and relations without dependency trees［C］. Vancouver：The 55th Annual Meeting of the Association for Computational Linguistics（ACL），2017.

7 ADEL H，SCHÜTZE H. Global normalization of convolutional neural networks for joint entity and relation classification［C］. Copenhagen：The 2017 Conference on Empirical Methods in Natural Language Processing（EMNLP），2017.

8 MIWA M，BANSAL M. End－to－end relation extraction using LSTMs on sequences and tree structures［C］. Berlin：The 54th Annual Meeting of the Association for Computational Linguistics（ACL），2016.

9 YU B，ZHANG Z，SHU X，et al. Joint extraction of entities and relations based on a novel decompositionstrategy［EB/OL］.［2021－12－10］. https：//www.researchgate.net/publication/335737705_Joint_Extraction_of_Entities_and_Relations_Based_on_a_Novel_Decomposition_Strategy.

10 ZENG X，ZENG D，HE S，et al. Extracting relational facts by an end－to－end neural model with copy mechanism［C］. Melbourne：The 56th Annual Meeting of the Association for Computational Linguistics（ACL），2018.

11 FU T J，LI P H，MA W Y. GraphRel：modeling text as relational graphs for joint entity and relation extraction［C］. Firenze：The 57th Annual Meeting of the Association for Computational Linguistics（ACL），2019.

12 CUI Y，CHE W，LIU T，et al. Pre－training with whole word masking for chinese bert［EB/OL］.［2021－12－10］. https：//www.researchgate.net/publication/333892280_Pre－Training_with_Whole_Word_Masking_for_Chinese_BERT.

13 GIANNIS B，JOHANNES D，THOMAS D，et al. Joint entity recognition and relation extraction as a multi－head selection problem［J］. Expert systems with application，2018，114（12）：34－45.