DOI:10.16781/j.0258-879x.2021.11.1273

·论　著·

基于缺血性脑卒中患者出院小结的协变量提取方法

林　振[1,2]，秦宇辰[1]，秦婴逸[1]，李冬冬[1]，吴　骋[1*]，贺　佳[1*]
1. 海军军医大学（第二军医大学）卫生勤务学系卫生统计学教研室，上海 200433
2. 中国人民解放军73127部队医院，福州 350000

［摘要］ **目的**　针对缺血性脑卒中这一发病率高、预后差的疾病，应用自然语言处理技术从患者出院小结中进行文本数据挖掘，并通过Python编程语言将非结构化的文本数据转换成供后续统计分析的结构化数据库。**方法**　利用缺血性脑卒中患者出院小结资料，构建基于知识增强的语义表示模型（ERNIE）＋神经网络＋条件随机场的命名实体识别模型，进行疾病、药物、手术、影像学检查、症状5种医疗命名实体的识别，提取实体构建半结构化数据库。为了进一步从半结构化数据库中提取出结构化数据，构建基于ERNIE的孪生文本相似度匹配模型，评价指标为准确率，采用最优模型构建协变量提取器。**结果**　命名实体识别模型总体F1值为90.27%，其中疾病F1值为88.41%，药物F1值为91.03%，影像学检查F1值为87.71%，手术F1值为87.07%，症状F1值为96.59%。文本相似度匹配模型的总体准确率为99.11%。**结论**　通过自然语言处理技术，实现了从完全的非结构化数据到半结构化数据再到结构化数据的构建流程，与人工阅读病历并手动提取病历信息相比，极大提高了数据库构建的效率。

［关键词］ 卒中；电子健康病历；患者出院小结；自然语言处理；命名实体识别；相似度匹配；协变量
［中图分类号］ R 197.324　［文献标志码］ A　［文章编号］ 0258-879X(2021)11-1273-06

Covariate extraction method based on discharge summary of stroke patients

LIN Zhen[1,2], QIN Yu-chen[1], QIN Ying-yi[1], LI Dong-dong[1], WU Cheng[1*], HE Jia[1*]
1. Department of Health Statistics, Faculty of Health Services, Naval Medical University (Second Military Medical University), Shanghai 200433, China
2. No. 73127 Troop Hospital of PLA, Fuzhou 350000, Fujian, China

［**Abstract**］ **Objective**　To carry out text data mining from discharge summary of patients with stroke (a disease with high incidence and poor prognosis) using natural language processing technology, and to convert unstructured text data into structured database for subsequent statistical analysis through Python. **Methods**　Based on the discharge summary of patients with ischemic stroke, the named entity recognition model of enhanced representation from knowledge integration (ERNIE)＋neural network＋conditional random field was constructed to identify 5 kinds of medical named entities, including disease, drug, surgery, imaging examination and symptoms. The entities were extracted and the semi-structured database was constructed. In order to further extract structured data from semi-structured databases, a similarity matching model of twin texts based on ERNIE was constructed. The evaluation index was accuracy, and the optimal model was used to construct the covariable extractor. **Results**　The overall F1 value of the named entity recognition model reached 90.27%, including 88.41% for disease F1, 91.03% for drug F1, 87.71% for imaging examination F1, 87.07% for surgery F1, and 96.59% for symptom F1. The overall accuracy of the text similarity matching model reached 99.11%. **Conclusion**　The construction process from complete unstructured data, to semi-structured data, and then to structured data, is realized through natural language processing technology. Compared with reading and extracting medical records manually, the natural language processing technology greatly improved the efficiency of database construction.

［**Key words**］ stroke; electronic health record; patient discharge summary; natural language processing; named entity recognition; similarity matching; covariate

［Acad J Sec Mil Med Univ, 2021, 42(11): 1273-1278］

［收稿日期］ 2021-05-21　［接受日期］ 2021-06-28
［基金项目］ 全军后勤科研重大项目子课题（AWS14R013-1），上海市公共卫生体系建设三年行动计划（2020－2022年）优秀人才培养计划（GWV-10.1-XD05）. Supported by Major Logistics Research Project of PLA (AWS14R013-1) and Outstanding Talent Trainning Plan of Shanghai 3-Year Action Plan (2020-2022) for Public Health System (GWV-10.1-XD05).
［作者简介］ 林　振，硕士生. E-mail: 3090100460@zju.edu.cn
*通信作者（Corresponding authors）. Tel: 021-81871442, E-mail: wucheng_wu@126.com; Tel: 021-81871441, E-mail: hejia63@yeah.com

电子病历是真实世界大数据中质量较高的部分。电子病历从21世纪初开始兴起，其使用率在2008年仅为9%，而在2015年已经上升到96%[1]。国务院于2016年发布的指导意见指出，应当将医疗大数据作为重要的战略性资源，通过有效的利用，激发医疗体制改革的动力[2]。电子病历包含了结构化数据、半结构化数据和非结构化数据。通常很多重要的临床信息都被记录为非结构化的文本，医生花费了大量时间来记录现病史、体格检查、病程记录等信息，这部分信息所占比重大（据专家估计这部分信息占总量的80%以上[3]），但利用率低，无法直接用于传统的统计分析。协变量是指可能会对主要变量分析产生影响的因素[4]，只有从文本中提取出相关的协变量并构建结构化数据库才能进行后续的统计分析。国外对于电子病历的非结构化数据研究较早，英文电子病历的分析和数据提取方法也较为成熟[5-7]。中文表述与英文相差较大，因此通过一系列技术手段将中文电子病历包含的非结构化数据构建成高质量的结构化数据库，意义重大且充满挑战。

近年来，自然语言处理（natural language processing）技术已经被广泛应用于从非结构化电子病历的信息提取过程，包括医疗命名实体识别（named entity recognition）、文本分类、共现分析等[8]。运用自然语言处理技术将非结构化的文本转换为结构化数据能够有效减少人工阅读文本提取数据的时间，提高了非结构化数据的可用性，从而实现大规模文本的自动处理。

本研究针对缺血性脑卒中这一发病率高、预后差疾病的患者出院小结文本资料，构建了一系列自然语言处理技术方法。首先利用基于深度学习的命名实体识别模型进行疾病、药物、手术、影像学检查、症状5种医疗命名实体的识别，提取实体后构建半结构化数据库，进而建立文本相似度匹配模型从半结构化数据库中提取出结构化数据，最终将非结构化文本资料转换成可供后续统计分析的结构化数据库。

1　资料和方法

1.1　资料来源　研究数据来源于上海市某三甲医院2009－2019年缺血性脑卒中患者的出院小结，共6 053例。所有与患者隐私相关的信息，如姓名、家庭住址等在获取数据前已由数据提供者去除。

1.2　数据预处理　出院小结为文本数据，包含了患者的疾病信息、用药信息、各项检查信息等，通过采用正则表达式“re.sub”函数，用“re.sub(r‘[^,。]*(未|否认|正常|(－)|阴性)[^,。]*。?’，“”，text)”语句将文本中包含否定词及阴性词的文本去除，最大程度保留阳性特征文本，同时采用正则表达式删除出院医嘱信息。

1.3　命名实体识别

1.3.1　命名实体标注　本研究中，主要将医疗命名实体识别聚焦在疾病、药物、手术、影像学检查、症状5个方面。数据采用BIO标注体系，B代表一个实体的起始，I代表一个实体除起始以外的后续部分，O代表非实体部分。B和I后面将会跟随该实体所属类别，以“小明突发脑卒中，感觉恶心、呕吐，在进行MRI检查后行动脉取栓术，开始服用阿司匹林”为例，如图1。随机选取1 000份缺血性脑卒中患者的出院小结进行标注，标注完成后由医院病案室专家进行核查。数据标注完成后将数据集按3 ： 1 ： 1的比例随机分成3份，分别为训练集、验证集和测试集。

小	明	突	发	脑	卒	中	，
O	O	O	O	B-疾病	I-疾病	I-疾病	O

感	觉	恶	心	、	呕	吐	，
O	O	B-症状	I-症状	O	B-症状	I-症状	O

在	进	行	M	R	I	检	查
O	O	O	B-影像学	I-影像学	I-影像学	O	O

后	行	动	脉	取	栓	术	，
O	O	B-手术	I-手术	I-手术	I-手术	I-手术	O

开	始	服	用	阿	司	匹	林
O	O	O	O	B-药物	I-药物	I-药物	I-药物

图1　命名实体标注示意图

Fig 1　Schematic diagram of named entity annotation

1.3.2　命名实体识别模型构建　命名实体识别模型的基本框架为预训练字嵌入＋神经网络＋条件随机场，即首先将文本信息通过预训练模型转换为字向量，随后输入到神经网络中，神经网络输出后再输入到条件随机场得到每个字的实体类别。本研究的预训练字向量、神经网络分别通过基于知识

增强的语义表示模型（enhanced representation from knowledge integration，ERNIE）和膨胀卷积神经网络（iterated dilated convolutional neural network，IDCNN）[9]，利用 Python 编程语言实现。

ERNIE 通过知识增强，利用先验的语义知识学习文本间的真实语义关系[10]。ERNIE 的模型结构与基于 Transformer 的双向编码表示模型（bidirectional encoder representations from transformer，BERT）相同，由输入层、基于双向 Transformer 的编码层及基于具体任务的输出层构成。与 BERT 遮盖字的方式不同，ERNIE 在进行掩码语言模型（masked language model，MLM）训练时采用 3 种不同的遮盖模式：第 1 种模式与 BERT 相同，随机抽取 15% 的字进行遮盖；第 2 种模式通过分词获得中文短语，随机抽取部分短语进行遮盖；第 3 种模式根据先验知识选取语料库中的人名、地名等实体随机遮盖。用来训练的语料库更多采用优质的中文语料库，如百度百科、中文维基百科等。

IDCNN 如图 2 所示，图 2A 为传统的 2 层 3×3 卷积核，感受野为 5，即第 i 层能感受到的上下文距离为 $2i+1$；图 2B 为 2 层的膨胀系数为 2 的卷积核，感受野为 7，即第 i 层能感受到的上下文距离为 $2^{i+1}-1$，可见传统的卷积核与上下文距离呈线性相关，而膨胀的卷积核与上下文距离呈指数相关，使神经网络能够捕捉到长距离的文本关系。

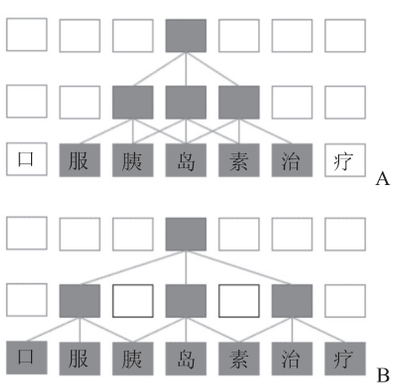

图 2　基于文本的膨胀卷积神经网络模型

Fig 2　Iterated dilated convolutional neural network model for text

A: Traditional 2-layer 3×3 convolution kernel; B: Two-layer iterated dilated convolution kernel with expansion coefficient of 2.

条件随机场是一类无向图模型[11]，在命名实体识别中最常用的是一种线性链结构，用来进行序列标签分析。对于给定的文字序列 $x=\{x_1, x_2, \cdots, x_n\}$，$x_i$ 表示第 i 个字符的特征向量，给定 x 对应的标签序列 $y=\{y_1, y_2, \cdots, y_n\}$，则有条件概率分布 $P(y|x)$：

$$P(y|x)=\frac{\sum_{t=1}^{n}\exp(S(y^t, x^t, \theta))}{\sum_{t=1}^{n}\sum_{y\in\gamma(x)}\exp(S(y_i, x^t, \theta))}$$

其中 $\gamma(x)$ 表示 x 的可能的标签，S 表示势函数，θ 为模型的参数，t 表示某时刻。根据维特比算法，利用条件随机场可以学习到上下文标签之间的关系，对输入的文字序列 $x=\{x_1, x_2, \cdots, x_n\}$ 求得全局最优的标签序列。

模型实验设置方面，ERNIE 预训练模型为 12 层 Transformer，多头注意力机制为 12 头，采用微调模式得到 768 维字向量，学习率为 5×10^{-5}，丢弃率值为 0.5，梯度截断值为 5，迭代次数 100 次。

1.4　协变量提取器

1.4.1　*建模数据集的构建*　在通过 1.3 节命名实体识别后得到的半结构数据集基础上，随机抽取了 1 000 例患者的命名实体，标注完成后由病案室专家进行核查。标注示例如图 3 所示，第 1 列为数据编号，第 2 列为实体类别，第 3 列为病历中抽取的医疗实体（实体 1），第 4 列为研究人员进行标注的标准实体（实体 2），第 5 列表示实体 1 与实体 2 是否匹配。为了获得建模数据集，在正样本数据库的基础上，采用数据增广技术进行负样本构建和正样本扩充。由于抽取的实体所对应的标准名称具有唯一性，因此将“实体 2”这一列打乱进行随机配对，当“实体 2”不再是原来的实体时，标记为“不匹配”。正样本的扩充采用相似性传递的方法，即实体 A 与实体 B 匹配，实体 C 与实体 B 匹配，则实体 A 与实体 C 匹配。最后将标注的数据集、构建的负样本数据集及扩充的正样本数据集合并，将数据集随机分为训练集、验证集、测试集，比例为 3 ∶ 1 ∶ 1。

1.4.2　*文本相似度匹配模型的构建*　基于 ERNIE 的模型如图 4 所示，利用孪生网络结构，首先将 2 个实体送入 ERNIE，ERNIE 的参数对 2 个实体共享，得到 2 个实体的句向量，随后送入汇聚层。采用平均汇聚方式对句向量进行特征提取和压缩，得到向量 u 和 v，最后将 u、v、$|u-v|$ 拼接后送入全连接层，通过 logistic 函数判断 2 个实体

是否相似。模型实验设置方面，ERNIE 采用 12 层 Transformer，隐藏层大小为 768，多头注意力机制为 12 头，优化器为 Adam，设置学习率为 2×10^{-5}，批量大小为 32，训练迭代 10 次。

No.	实体类别	医疗实体(实体 1)	标准实体(实体 2)	是否匹配
1	疾病	脑梗死(大动脉粥样硬化型)	脑梗塞	是
2	疾病	急性左心衰竭	心力衰竭	是
3	疾病	风心病	风湿性心脏病	是
4	疾病	腔隙性脑梗死	脑梗塞	是
5	疾病	慢性胃溃疡(稳定期)	胃溃疡	是
6	手术	垂体瘤切除	垂体瘤切除术	是
7	手术	右侧侧窦区脑动静脉瘘栓塞术	动静脉瘘栓塞术	是
8	手术	左侧颈动脉取栓术	动脉取栓术	是
9	手术	颈内动脉支架成形术	动脉支架成形术	是
10	手术	颅骨钻孔颅内外血流重建术	颅内外血运重建术	是
11	药物	阿司匹林片	阿司匹林	是
12	药物	阿托伐他汀钙片	阿托伐他汀	是
13	药物	氨溴索注射液	氨溴索	是
14	药物	盐酸二甲双胍	二甲双胍	是
15	药物	氨氯地平(络活喜)片	氨氯地平	是
16	影像学检查	腹部B超	超声	是
17	影像学检查	颈动脉CT增强造影	CT	是
18	影像学检查	左侧颈内动脉MRI	MRI	是
19	影像学检查	电子胃镜	胃镜	是
20	影像学检查	腹主动脉CTA	CTA	是
21	症状	肢体不灵活	肢体功能障碍	是
22	症状	吞咽受限	吞咽障碍	是
23	症状	认知障碍	认知功能障碍	是
24	症状	言语构音不清	构音障碍	是
25	症状	反应稍迟钝	反应迟钝	是

图 3　文本相似度匹配正样本示例

Fig 3　Positive sample example of text similarity matching model

CT: Computed tomography; MRI: Magnetic resonance imaging; CTA: Computed tomography angiography.

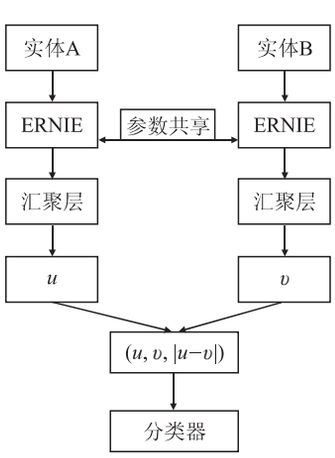

图 4　ERNIE 文本相似度匹配模型

Fig 4　ERNIE text similarity matching model

ERNIE: Enhanced representation from knowledge integration.

1.5　*研究平台*　本研究平台由海军军医大学（第二军医大学）卫生勤务学系军队卫生统计学教研室及长海医院信息科提供，服务器采用 Windows Server 2008 R2 64 位操作系统，8 核 Intel（R）Xeon（R）CPU。软件环境采用 Python 3.7 及 TensorFlow 1.10。

2　结　果

2.1　*命名实体识别结果*　命名实体识别模型的评价指标采用精确率、召回率和 F1 值。精确率＝真阳性 /（真阳性＋假阳性），即阳性预测值，表示在预测为阳性的样本中真正是阳性的样本所占的比例。召回率＝真阳性 /（真阳性＋假阴性），即灵敏度，表示所有的阳性样本中正确预测为阳性的比例。F1 值＝（2× 精确率 × 召回率）/（精确率＋召回率），F1 值综合考虑了精确率和召回率，

是精确率和召回率的调和平均。实验结果显示，模型总体精确率为 91.07%，其中疾病、药物、影像学检查、手术、症状的精确率分别为 90.78%、91.79%、87.55%、85.59%、95.74%；模型总体召回率为 89.49%，其中疾病、药物、影像学检查、手术、症状的召回率分别为 86.17%、90.29%、87.87%、88.60%、97.46%；模型总体 F1 值达到了 90.27%，其中疾病、药物、影像学检查、手术、症状的 F1 值分别为 88.41%、91.03%、87.71%、87.07%、96.59%。

2.2　文本相似度匹配结果　文本相似度匹配模型的评价指标为准确率，准确率（%）＝预测正确实体对数目 / 总实体对数目 ×100%。ERNIE 的总体准确率达到了 99.11%，其中疾病、药物、影像学检查、手术、症状的准确率分别为 99.54%、99.00%、98.40%、98.61%、97.49%。

2.3　协变量提取器输出结果　通过上述步骤，研究者可根据需要提取的协变量，在协变量提取器中输入相应的医疗实体后，即可得到如图 5 所示的结构化数据库。第 1 列为患者编号，其后每列表示某种疾病、手术、症状或药物，如列名所示，表中数字表示该列所示实体是否发生，“1” 表示发生，“0” 表示未发生。例如，患者 1 仅动脉支架形成术赋值为 “1”，表示该患者接受了这一操作，且该患者无高血压、高脂血症、糖尿病等所列疾病，也未接受气管插管、CT 血管造影、CT 灌注成像等操作。

No.	高血压	高脂血症	糖尿病	心房颤动	动脉支架成形术	气管插管	CTA	CTP	失语	言语障碍	面舌瘫	肝素	疏血通
1	0	0	0	0	1	0	0	0	0	0	0	0	0
2	1	1	1	0	0	0	1	1	0	0	0	0	1
3	0	0	0	1	0	0	0	0	1	1	0	0	0
4	0	0	1	0	0	0	1	1	0	1	0	1	0
5	1	0	0	0	0	0	1	1	0	1	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	1	0	0	0	0	1	0	0	0
8	1	0	0	0	0	0	0	0	0	1	0	0	0
9	1	0	0	1	0	0	0	0	0	1	1	0	1
10	1	0	0	0	1	0	0	0	0	0	0	0	0
11	1	0	1	0	1	0	0	0	0	1	0	0	0
12	1	0	0	1	0	0	1	1	0	1	0	1	0
13	1	0	0	1	0	0	1	1	0	1	0	0	0
14	1	0	1	0	0	0	0	0	0	0	0	0	0
15	1	0	1	0	1	0	0	0	0	0	0	0	0
16	0	0	0	0	1	0	0	0	0	1	0	0	0
17	1	1	1	0	1	0	1	0	0	0	0	0	0
18	1	0	0	0	0	0	1	1	0	1	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	1	0	0	0	0	0	0	1	0	0	0
21	1	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	1	0	0	0
23	1	0	0	0	0	0	0	0	0	1	0	0	0
24	1	0	0	0	0	0	0	0	0	1	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	1	0	0	0
27	0	0	0	0	0	0	0	1	0	1	0	0	0
28	0	0	0	0	0	0	0	0	0	1	0	0	0
29	0	0	0	0	1	0	0	0	0	1	0	0	0
30	1	0	0	0	0	0	1	1	0	1	0	0	0

图 5　结构化数据集结果示例

Fig 5　An example of structured dataset results

In columns 2-14, “0” indicates none and “1” indicates yes. CTA: Computed tomography angiography; CTP: Computed tomography perfusion.

3 讨　论

本研究聚焦缺血性脑卒中患者的出院小结文本资料，包含了患者的诊断、症状、治疗等信息。通过构建“预训练模型＋神经网络模型＋条件随机场”的架构识别出院小结中的医疗命名实体，并利用此模型构建了一个包含疾病、手术、药物、影像学检查、症状的半结构化数据库；通过文本相似度匹配技术，开发出一个协变量提取器，可以在半结构化数据库中直接提取出满足分析需要的结构化数据，实现了从非结构化数据到半结构化数据再到结构化数据的构建流程。与人工阅读病历、手动提取病历信息相比，本研究所采用的方法可极大提高循证数据库的构建效率。

本研究通过预训练模型及神经网络模型，对医疗命名实体及其结构化数据提取进行了研究，取得了较好的实验结果。以往对于命名实体识别的研究集中在人名、地名、机构名等方面[12-14]，对于医疗实体的研究较少。医疗实体具有独特特征，分类较多，同一医疗实体的表述众多，无法通过编写词典库穷尽，因此需要通过深入挖掘上下文之间的关系来找出特定的实体。而深度学习能够通过学习到医疗文本深层次的隐含特征来进行命名实体的识别。本研究中，我们通过预训练模型将文本向量化。ERNIE通过微调，可以自行根据上下文的不同来调整字向量，能够更好地表达其在具体语境中的含义，解决了一词多义的问题，使得命名实体识别模型的效果得到了提升。

对于实体识别后提取出的各类医疗实体，由于同一实体可能有多种说法且难以统一规范，仍不满足进行统计分析的需求，因此本文提出构建一个基于文本相似度匹配的协变量提取器，在所构建的包含医疗实体的半结构化数据库基础上，通过相似度匹配，自动查找某一个体是否包含这一医疗实体，用“0”表示不存在该实体，“1”表示存在该实体，从而完成结构化数据库的构建。传统的文本相似度识别模型首先计算出相似度，然后通过设定阈值或排序来确定文本是否匹配，这种方法往往受人为因素干扰，阈值大小的设定对结果影响很大，而本研究通过有监督学习的方法利用文本相似度匹配技术来实现实体的统一，能够较为精确地提取出所需的协变量。

［参考文献］

［1］ KIM E, RUBINSTEIN S M, NEAD K T, WOJCIESZYNSKI A P, GABRIEL P E, WARNER J L. The evolving use of electronic health records (EHR) for research[J]. Semin Radiat Oncol, 2019, 29: 354-361.

［2］ 国务院办公厅. 国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见[EB/OL].(2016-06-24)[2021-05-20]. *http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.*

［3］ MARTIN-SANCHEZ F, MARTIN-SANCHEZ F, VERSPOOR K. Big data in medicine is driving big changes[J]. Yearb Med Inform, 2014, 9: 14-20.

［4］ 黄丽红，陈峰. 临床试验中协变量的处理[J]. 中国循证医学杂志，2019，19：1498-1502.

［5］ RAZJOUYAN J, FREYTAG J, DINDO L, KIEFER L, ODOM E, HALASZYNSKI J, et al. Measuring adoption of patient priorities-aligned care using natural language processing of electronic health records: development and validation of the model[J/OL]. JMIR Med Informatics, 2021, 9: e18756. DOI: 10.1093/geroni/igaa057.592.

［6］ WEEGAR R. Applying natural language processing to electronic medical records for estimating healthy life expectancy[J/OL]. Lancet Reg Health West Pac, 2021, 9: 100132. DOI: 10.1016/j.lanwpc.2021.100132.

［7］ ANNAPRAGADA A V, DONARUMA-KWOH M M, ANNAPRAGADA A V, STAROSOLSKI Z A. A natural language processing and deep learning approach to identify child abuse from pediatric electronic medical records[J/OL]. PLoS One, 2021, 16: e0247404. DOI: 10.1371/journal.pone.0247404.

［8］ JUHN Y, LIU H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research[J]. J Allergy Clin Immunol, 2020, 145: 463-469.

［9］ YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. arXiv:1511.07122[cs.CV]. (2016-04-30)[2021-05-20]. *https://arxiv.org/abs/1511.07122.*

［10］ SUN Y, WANG S H, LI Y K, FENG S K, CHEN X Y, ZHANG H, et al. ERNIE: enhanced representation through knowledge integration[EB/OL]. arXiv:1904.09223[cs.CL]. (2019-04-19)[2021-05-10]. *https://arxiv.org/abs/1904.09223.*

［11］ 余本功，范招娣. 面向自然语言处理的条件随机场模型研究综述[J]. 信息资源管理学报，2020，10：96-111.

［12］ 胡新棒，于溆乔，李邵梅，张建朋. 基于知识增强的中文命名实体识别[J/OL]. 计算机工程，2021. DOI：10.19678/j.issn.1000-3428.0059810.

［13］ 黄晓辉，乔立升，余文涛，李京，薛寒. 中文分词与命名实体识别的联合学习[J]. 国防科技大学学报，2021，43：86-94.

［14］ 马孟铖，杨晴雯，艾斯卡尔•艾木都拉，吐尔地•托合提. 基于词向量和条件随机场的中文命名实体分类[J]. 计算机工程与设计，2020，41：2515-2522.

［本文编辑］　杨亚红