信息抽取在构建医学知识图谱中的应用及进展*

匡泽民 李健铨 邓 楠

（首都医科大学附属北京安贞医院 北京 100029） （北京神州泰岳软件股份有限公司 北京 100020）

〔摘要〕 阐述医学信息抽取中实体识别、实体消歧和关系抽取 3 个重要步骤，介绍传统方法、基于机器学习以及基于深度学习的应用，对前沿内容和未来发展进行展望。

〔关键词〕 医学知识图谱；关系抽取；机器学习；深度学习

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.01.007

Application and Progress of Information Extraction in Building Medical Knowledge Graph *KUANG Zemin, Beijing Anzhen Hospital of Capital Medical University, Beijing 100029, China; LI Jianquan, DENG Nan, Beijing Ultrapower Software Company Ltd., Beijing 100020, China*

〔**Abstract**〕 The paper expounds three important steps of medical information extraction, including entity recognition, entity disambiguation and relation extraction, introduces the application of traditional method, machine learning and deep learning, and looks forward to the frontier content and future development.

〔**Keywords**〕 medical knowledge graph; relation extraction; machine learning; deep learning

1 引言

知识图谱是由实体作为节点组成的巨大语义网络，实体之间关系或属性形成网络边界。作为一种高效知识管理方式知识图谱可以支撑知识检索、问答、可视化分析等任务场景。构建医学知识图谱的关键在于如何正确高效提取知识中实体间的关系，即信息抽取。而医学领域关系抽取标准较其他领域更严格，同时实体间相互关联导致实体关系更复杂。因此借助人工智能（Artificial Intelligence，AI）技术构建医学知识图谱，实现医学实体间关系分析与挖掘对临床诊疗工作具有重要意义。医学信息抽取核心任务是实体识别、实体消岐和关系抽取。本文拟从上述 3 个子任务入手，描述利用传统方法、机器学习方法和深度学习方法识别、消岐实体以及明确实体间关系的相关 AI 技术。

〔收稿日期〕 2020－05－18

〔作者简介〕 匡泽民，博士，主任医师，硕士生导师，发表论文 50 余篇，获软件著作权 5 项。

〔基金项目〕 中国中青年临床研究基金－VG 研究基金“未达标高血压患者基于互联网精准管理效果的多中心随机对照研究”（项目编号：2017－CCA－VG－016）；北京安贞医院院长发展基金项目“老年人群高血压‘四维监测’管理新策略及效果评价”（项目编号：2016－P－01）。

2 实体识别

2.1 概述

医学知识信息抽取第1步是识别文本中的实体和关键概念，即实体识别。实体类型包括患者信息、药品属性信息、医学概念、时间表达式等，其中医学概念识别是医学信息抽取研究核心问题。以“腹平坦未见腹壁静脉曲张”为例，“腹”与“腹部”都被定为“body”即身体部分，而“静脉曲张”被识别为“symptom”即症状，见图1。

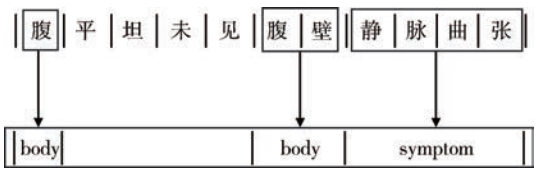

图1 实体识别样例

2.2 基于规则的方法

基于规则识别是传统实体识别中最早使用的办法。用于生物医学信息提取的规则一般依靠医学知识资源，通过人工总结或自然语言处理方法构建模板。基于规则识别具有及时、高效、不需要人为标注数据等优点。使用的规则可以来自构建好的字典或专家系统，比如Khalifa等[1]使用cTAKES标记好的文本进行分配，通过统一医学语言系统（Unified Medical Language System，UMLS）模块的字典索引识别疾病和风险因素项目。常用的自然语言处理（Natural Language Processing，NLP）系统还包括MedLEE、KnowledgeMap和MetaMap等。但由于自然语言表达的复杂性与多样性，现有规则和模版并不能涵盖所有可能出现的情况。在已有字典或系统不足以解决当前问题的情况下也可使用迭代方法进行自定义规则创建，每次迭代都对规则结果进行一次评估，如果出现问题则由专家手动修改直到所有问题都解决[2-4]。

2.3 基于机器学习的方法

基于机器学习识别实体信息是将医学文本中识别实体任务视为分类问题。机器学习算法将定义好的特征传递给分类器并对目标实体进行识别。常用模型包括支持向量机（Support Vector Machine，SVM）模型和条件随机场（Conditional Random Fields，CRF）模型等。Singh等使用SVM提取特征对糖尿病患者中高血压患病进行诊断[5]，构造出一种新特征选择机制，产生可理解的规则集；再用XGBoost作为决策工具将训练集合传入并得到分类规则。实验表明该方法优于此前提出的多种基准分类器。CRF将实体识别问题转换为序列标注问题[6]，Liao等[7]针对数据集对CRF框架模型进行优化，使得模型考虑到距离更远的上下文关系，最终性能达73.20%。Lee等[8]在外部子序列隐藏变量中增加记忆元素，引导信息传递、降低远距离标签计算时成本。混合SVM和CRF优点的结构化支持向量机（Structured Support Vector Machine，SSVMs）在SVM基础上引入结构化信息且取得超越CRF的效果[9]。机器学习方法可一定程度实现信息抽取自动化，但对特征工程的人工选择结果依赖性比较大且容易产生衍生错误。目前基于机器学习的实体识别在小规模数据集取得可靠效果，在大数据集上效果尚待改善。

2.4 基于深度学习的方法

基于深度学习的方法在实体识别领域有广泛应用，最常用的是基于双向长短期记忆网络（Bidirectional Long Short-Term Memory Network，BiLSTM）CRF，即BiLSTM-CRF模型[10]，包含输入层、隐藏层和输出层3层结果，解决长遗忘问题，输出计算全局最优解。杨培等[11]在该模型基础上引入注意力（Attention）机制获取所关注词上下文表示，该方法F值最高可达90.77%，极大提高实体识别效果。Wang Q等[12]使用5种不同特征表达机制处理中文临床实体识别任务，对BiLSTM-CRF模型进行扩展，将data-driven深度学习方法和knowledge-driven词典方法结合，把识别任务当作句子标注任务因而获得较好的模型表现。而BERT模型[13]使得实体识别技术更进一步，Alsentzer等[14]在领域特定模型上引入BERT，训练两个BERT变种，用于

临床文本的 $BERT_{BASE}$ 称为临床 BERT，用于出院指南的 BioBERT[15] 称为出院总结 BERT，指出使用特定模型能明显提升性能。Yuqi Si 等[16] 比较 $BERT_{BASE}$ 和 $BERT_{LARGE}$ 在临床概念识别上的应用，其中 $BERT_{LARGE}$ 在 i2b2 2010 上 F1 分数达 90.25%，在 i2b2 2012 上 F1 分数达 80.91%，超过传统嵌入模型表现。深度学习模型更容易得到复杂特征，脱离特征工程约束，自主从输入中发掘和学习信息，但是深度学习模型一般需要大量标注语料，可选网络种类繁多、可解释性差。深度学习模型各有优点，在实体识别上的应用仍有很大发展空间。实体识别方法总结，见图 2。

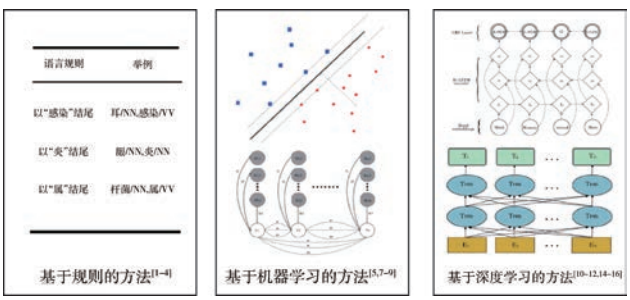

图 2 实体识别

3 实体消岐

3.1 概述

医学实体存在简写、缩写、不规范、一词多义等问题，可分为 3 种类别：知识缺失、多样性问题和歧义性问题，迫切需要通过医学实体消岐解决。中文语法的简洁性容易导致语义混淆。以中文医学实体为例，“近端骨折”和“上端骨折”都与实体“上端骨折”相匹配，这意味着其含义一致，见图 3。为解决此类问题一般在进行实体关系抽取之前需要对实体进行消岐处理。实体消岐是为实体确定唯一标识符，建立实体与唯一标识符之间的映射。

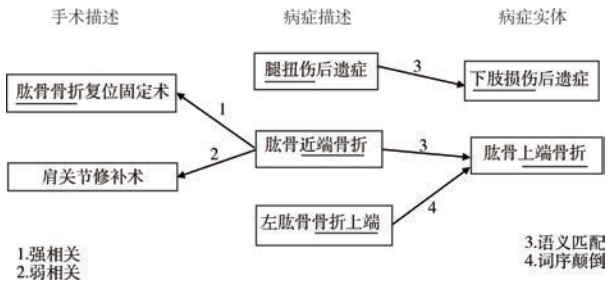

图 3 医学实体问题样例

3.2 基于查询与匹配的方法

大部分传统消岐工作基于查询或字符串匹配实现。为实现识别实体到知识库实体之间的映射需要计算实体指称项与目标实体之间的相似度，基于相似度对目标实体进行选择，指向同一目标实体的指称项都被判定为同一类。Han 等[17] 使用向量空间模型计算指称项和实体间相似性，将其作为图关系算法的一部分，通过图实现全局最优结果。胡运翠等[18] 利用语义相似度对识别到的基因实体进行消歧，实验结果准确率达到 80%，证明该方法适用于生物医学领域。为了充分利用有限资源通常使用字符串精确和模糊匹配相结合方法，精确匹配要求匹配的两个字符完全一致，但允许忽略大小写和连字符，而模糊匹配则弥补精确匹配结果中的遗漏。模糊匹配为词典中的项建立索引，然后把识别出来的实体当作查询项去索引中进行检测。BM25 是常用检测算法，其把词典实体看作文档，把目标集合中实体当作查询，按照 BM25 公式对分支进行计算，将得分超过一定阈值的实体作为匹配结果。传统消歧方法有助于改良实体识别结果，但容易忽视识别到的指称项之间的关系，匹配方法也容易建立错误的实体到指称项的映射。

3.3 基于机器学习的方法

近年来机器学习与深度学习方法逐渐应用于实

体消岐任务。机器学习方法常用模型包括逻辑回归、马尔可夫模型等。Tsuruoka[19]等使用逻辑回归方法计算蛋白质实体和词典中实体的相似度并在不同知识库上检验该方法，取得超越大部分模型的结果。Leaman 等[20]采用半马尔可夫链模型联合消岐和识别过程，同时提高两个过程的性能，该方法在疾病知识库中 F－score 为 0.807，在化学知识库中高达 0.895。

3.4 基于深度学习的方法

深度学习在实体消岐中应用不多，典型方法是使用 BiLSTM 来匹配单词上下文和实体语义[21－22]，该模型有效提高消岐算法性能。Ji Z 等[23]引入 BERT 中的双向编码器表示引用，在 3 种数据集上进行训练，把候选概念排序问题当作一个句子对分类任务，把生成的候选和原实体作为输入，对预训练好的多种 BERT 模型进行微调，准确率提高 1.17%，超越目前最先进的生物医学实体消岐方法。Li F 等[24]探究训练数据域对基于 BERT 的模型影响，使用 150 万未标记电子健康记录对 BioBERT 模型进行微调，后又在药物用法、药物适应症、药物不良反应 3 个语料库进一步训练，证明基于 BERT 模型可用于各种类型实体消岐任务。Ishani Mondal[25]提出基于 Triplet Network 的实体消岐框架，模型分为候选集生成和排序两部分，使用卷积神经网络（Convolutional Neural Network，CNN）提取特征后计算提及正负样本对距离，在计算损失时引入合页损失函数，不仅实现了较高准确率还解决了缩写问题。深度学习方法解决了传统方法无法处理的概念间内在联系问题，可以自动学习目标实体和文档表示之间的联合，但是其结果容易受上下文建模影响，效果仍有提升空间。利用不同方法进行实体消岐，见图 4。

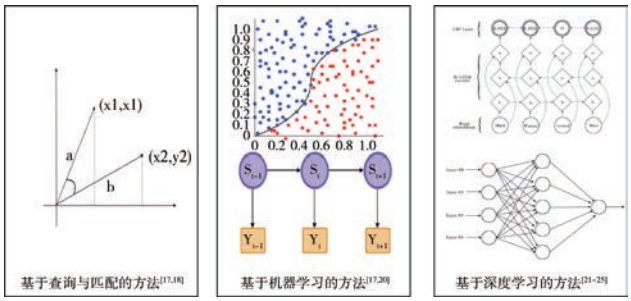

图 4 实体消岐

4 关系抽取

4.1 概述

实体关系抽取主要目的是抽取出实体之间对应关系，在生物医学领域中，这些对应关系通常包括基因与药物、基因与疾病、疾病与治疗关系等。解决实体关系抽取问题的常用方法是将其视为实体之间关系的分类问题。

4.2 基于机器学习的方法

基于机器学习方法的关系抽取可分为基于特征和基于核两类。基于特征的抽取方法分两个阶段，第 1 阶段用标注数据训练多个分类器，第 2 阶段用分类器判断实体间关系。其常用方法是最大熵模型和 SVM[26－27]，这些方法都需要大量标注语料。因此对标注语料要求较少的半监督、弱监督和无监督机器学习方法成为不错的选择。Bootstrapping 方法[28]是常用的方法之一，用来解决人工标注语料问题。基于核抽取方法不关注实体类型特征，而是将实体关系编码为特定结构，如树、序列等。前者具有高效快速优点，但过分依赖特征选择过程且选择特征占用大量时间；后者不需构造固有特征空间且可以利用上下文结构特征，但抽取速度较慢。两种抽取方法可以相互补充。

4.3 基于深度学习的方法

在医学类关系抽取任务中即使只使用简单网络结构也能实现让人满意的效果。Zhang Q 等[29]使用深度学习方法和传统中医相结合探究高血压治疗方法，提出包含两个堆叠隐藏层的自编码模型，将输入编码至隐藏层再解码到输出层来完成特征学习，在中药理论基础上对用药情况进行分析。深度神经网络的常见模型卷积神经网络（CNNs）最初主要应用于图像类任务，Kim Y 等[30]将其引入文本和关系分类任务并提出专用于 NLP 处理的 CNNs 模型框架。目前 CNNs 在生物医药领域的应用逐步扩展[31-32]。以 *A Shortest Dependency Path Based Convolutional Neural Network for Protein – Protein Relation Extraction* [32]中提到的最短依赖路径算法为例，该算法对句子中每个词进行分析得出语法依赖结果，其中绿色文字是目标蛋白质，红色箭头代表蛋白质之间最短依赖路径，见图 5。在医学信息提取中 RNN 各种变体比如 LSTM 模型和 BERT 更常用，这些方法近年展露出不凡的竞争力[33-35]。Lin[36]利用 BERT 提取临床概念关系，可跨越多个句子从而实现跨行关系提取，行内和跨行关系提取结果熵都有所提升。Can Tian 等[37]探究将关系抽取任务转化为标记问题，通过端对端模型将实体和事件连接，显示其效果超越大部分模型，但存在无法处理同一句中多个事件的问题。北京协和医院 Zhang X 等[38]在无标记临床语料库上引入 BERT 模型，其表现超越目前最好的关系抽取模型。基于深度学习的关系抽取实现“自动抽取”，彻底解放手工定义特征过程，但是深度学习方法受限于预先设定好的关系集合，对于语料中未标记的关系神经网络无法自动发现新关系。应用不同方法的实体关系抽取，见图 6。

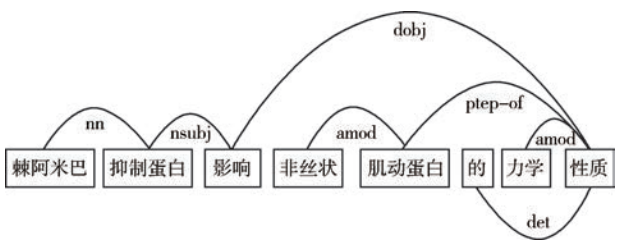

图 5 最短依赖路径抽取实体关系

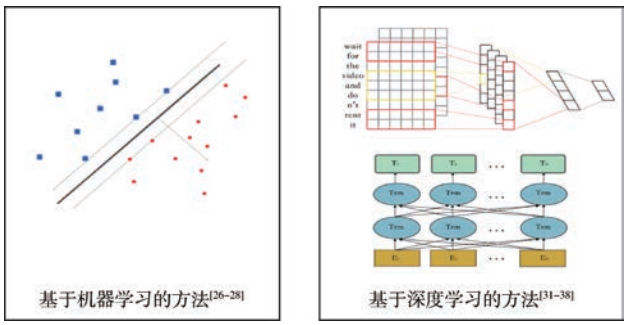

图 6 实体关系抽取

5 结语

信息抽取方法在医学领域有广泛应用，但医学文本信息的复杂性、多样性给医学信息抽取带来困难，使得各种抽取系统具有局限性。随着 AI 新技术的快速发展这一问题必将被克服，带动医学关系抽取技术水平提升，进而推动生物医学发展。

参考文 献

1 Khalifa A, Meystre S. Adapting Existing Natural Language Processing Resources for Cardiovascular Risk Factors Identification in Clinical Notes [J]. Journal of Biomedical Informatics, 2015 (58): S128 – S132.

2 James Cormack, Chinmoy Nath, David Milward, et al. Agile Text Mining for the 2014 i2b2/UTHealth Cardiac Risk Factors Challenge [J]. Journal of Biomedical Informatics, 2015 (58): S120 – S127.

3 Kelahan LC, Fong A, Ratwani R, et al. Call Case Dashboard: tracking R1 exposure to high – acuity cases using natural language processing [J]. Journal of the American College of Radiology, 2016, 13 (8): 988 – 991.

4 Jie W, Yan P, Xiaoxiao R, et al. An Expert System for Diagnosis and Treatment of Hypertension Based on Ontology [C]. Singapore: International Conference on Bio – Inspired Computing: Theories and Applications, 2018: 264 – 274.

5 Singh N, Singh P, Bhagat D. A Rule Extraction Approach from Support Vector Machines for Diagnosing Hypertension among Diabetics [J]. Expert Systems with Applications, 2019 (130): 188 – 205.

6 Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: probabilistic models for segmenting and labeling sequence data [C]. San Francisco: Proceedings of the Eighteenth International Conference on Machine Learning, 2001: 282 – 289.

7 Liao Z, Wu H. Biomedical Named Entity Recognition Based on Skip – chain Crfs [C]. Xi'an: 2012 International Conference on Industrial Control and Electronics Engineering. IEEE, 2012: 1495 – 1498.

8 Lee W, Choi J. Connecting Distant Entities with Induction through Conditional Random Fields for Named Entity Recognition: precursor – induced CRF [C]. Melbourne: Proceedings of the Seventh Named Entities Workshop, 2018: 9 – 13.

9 Tang B, Cao H, Wu Y, et al. Recognizing Clinical Entities in Hospital Discharge Summaries Using Structural Support Vector Machines with Word Representation Features [J]. Bmc Medical Informatics & Decision Making, 2013, 13 (S1): 1 – 10.

10 Huang Z, Xu W, Yu K. Bidirectional LSTM – CRF Models for Sequence Tagging [J]. Computation and Language, 2015, 3 (4): 1508 – 1991.

11 杨培, 杨志豪, 罗凌, 等. 基于注意机制的化学药物命名实体识别 [J]. 计算机研究与发展, 2018, 55 (7): 1548 – 1556.

12 Qi Wang, Yangming Zhou, Tong Ruan, et al. Incorporating Dictionaries into Deep Neural Networks for the Chinese Clinical Named Entity Recognition [J]. Journal of Biomedical Informatics, 2018 (92): 103133.

13 Devlin J, Chang M W, Lee K, et al. BERT: pre – training of deep bidirectional transformers for language understanding [C]. Minneapolis: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171 – 4186.

14 Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings [C]. Minneapolis: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019: 72 – 78.

15 Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT: a pre – trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2019, 36 (4): 1234 – 1240.

16 Yuqi Si, Jingqi Wang, Hua Xu, et al. Enhancing Clinical Concept Extraction with Contextual Embeddings [J]. Journal of the American Medical Informatics Association, 2019 (26): 1297 – 1304.

17 Han X, Sun L, Zhao J. Collective Entity Linking in Web Text: a graph – based method [C]. Beijing: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011: 765 – 774.

18 胡运翠, 林鸿飞, 杨志豪. 语义相似度的基因名标准化方法 [J]. 计算机工程与应用, 2011 (35): 128 – 131.

19 Yoshimasa Tsuruoka, John McNaught, Junichi Tsujii, et al. Learning String Similarity Measures for Gene Protein Name Dictionary Look – up Using Logistic Regression [J]. Bioinformatics (Oxford, England), 2007 (23): 2768 – 2774.

20 Leaman R, Z Lu. TaggerOne: joint named entity recognition and normalization with Semi – Markov models [J]. Bioinformatics, 2016 (32): 2839 – 2846.

21 Luo A, Gao S, Xu Y, et al. Deep Semantic Match Model for Entity Linking Using Knowledge Graph and Text [J]. Procedia Computer Science, 2018 (129): 110 – 114.

22 Hui Chen, Baogang Wei, Yonghuai Liu, et al. Bilinear Joint Learning of Word and Entity Embeddings for Entity Linking [J]. Neurocomputing, 2017 (294): 12 – 18.

23 Ji Z, Wei Q, Xu H. BERT – based Ranking for Biomedical Entity Normalization [EB/OL]. [2020 – 05 – 15]. https://www.researchgate.net/publication/341829911_BERT – based_Ranking_for_Biomedical_Entity_Normalization.

24 Li F, Jin Y, Liu W, et al. Fine – tuning Bidirectional Encoder Representations from Transformers (BERT) – based Models on Large – scale Electronic Health Record Notes: an empirical study [J]. JMIR Medical Informatics, 2019 (7): e14830.

25 Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, et al. Medical Entity Linking Using Triplet Network [C]. Minneapolis: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019: 95 – 100.

26 Zhang Z. Weakly – supervised Relation Classification for In-

formation Extraction [C]. Washington DC: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, 2004: 581 - 588.

27 Suchanek F M, Ifrim G, Weikum G. Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents [C]. Philadelphia: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006: 712 - 717.

28 Rozenfeld B, Feldman R. High - performance Unsupervised Relation Extraction from Large Corpora [C]. Hong Kong: Sixth International Conference on Data Mining (ICDM 06). IEEE, 2006: 1032 - 1037.

29 Qingchen Zhang, Changchuan Bai, Zhikui Chen, et al. Smart Chinese Medicine for Hypertension Treatment with a Deep Learning Model [J]. Journal of Network and Computer Applications, 2019 (129): 1 - 8.

30 Kim Y. Convolutional Neural Networks for Sentence Classification [C]. Doha: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1746 - 1751.

31 Liu S, Tang B, Chen Q, et al. Drug - drug Interaction Extraction via Convolutional Neural Networks [J]. Computational and Mathematical Methods in Medicine, 2016 (2016): 1 - 8.

32 Hua L, Quan C. A Shortest Dependency Path Based Convolutional Neural Network for Protein - protein Relation Extraction [J]. BioMed Research International, 2016 (2016): 1 - 9.

33 Pandey C, Ibrahim Z, Wu H, et al. Improving RNN with Attention and Embedding for Adverse Drug Reactions [C]. Beijing: Proceedings of the 2017 International Conference on Digital Health, 2017: 67 - 71.

34 Xu B, Shi X, Zhao Z, et al. Leveraging Biomedical Resources in Bi - LSTM for Drug - drug Interaction Extraction [J]. IEEE Access, 2018 (6): 33432 - 33439.

35 Qin Y, Zeng Y. Research of Clinical Named Entity Recognition Based on Bi - LSTM - CRF [J]. Journal of Shanghai Jiaotong University (Science), 2018 (23): 392 - 397.

36 Lin C, Miller T, Dligach D, et al. A BERT - based Universal Model for Both within - and Cross - sentence Clinical Temporal Relation Extraction [C]. Minneapolis: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019: 65 - 71.

37 Tian C, Zhao Y, Ren L. A Chinese Event Relation Extraction Model Based on Bert [C]. Chengdu: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, 2019: 271 - 276.

38 Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, et al. Extracting Comprehensive Clinical Information for Breast Cancer Using Deep Learning Methods [J]. International Journal of Medical Informatics, 2019 (132): 103985.

(上接第 15 页)

34 陈柯羽, 杜清, 张华, 等. 精准医疗背景下的药品研发策略 [J]. 转化医学电子杂志, 2018, 5 (10): 55 - 61.

35 龚兆龙, 林毅晖, 袁泰昌, 等. 精准医学时代的抗肿瘤药物研发 [J]. 药学进展, 2017, 41 (2): 97 - 100.

36 赵鹏, 马泽君, 乐嘉伟. 银行数据资产安全分级标准与安全管理体系建设方法 [C]. 北京: 软科学国际研讨会, 2012.

37 李松涛, 谢宗晓. 数据分类/分级及其相关标准解析 [J]. 中国质量与标准导报, 2019 (4): 14 - 16.

38 王敏. 大数据时代如何有效保护个人隐私? ——一种基于传播伦理的分级路径 [J]. 新闻与传播研究, 2018, 25 (11): 69 - 92, 127 - 128.

39 Haendel M A, Chute C G, Robinson P N. Classification, Ontology, and Precision Medicine [J]. New England Journal of Medicine, 2018, 379 (15): 1452 - 1462.

40 Lin E, Tsai S. Multi - omics and Machine Learning Applications in Precision Medicine [J]. Current Pharmacogenomics & Personalized Medicine, 2017, 15 (2): 97 - 104.

41 Tebani A, Afonso C, Marret S, et al. Omics - based Strategies in Precision Medicine: toward a paradigm shift in inborn errors of metabolism investigations [J]. International Journal of Molecular Sciences, 2016, 17 (9): 1555.