计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):200−205 [doi: 10.15888/j.cnki.csa.008046]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 GRU 和 PCNN 的电力知识抽取①

宋厚岩[1,2], 王汉军[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
通讯作者: 宋厚岩, E-mail: songhouyan@126.com

摘　要: 构造电力系统知识图谱最重要的就是电力知识的抽取, 针对目前传统基于监督学习、单一神经网络模型存在的问题和缺点, 如 CNN 擅长提取局部最重要特征而不适合处理序列输入; RNN 在处理序列化任务占优势却对于重要特征的提取很乏力, 因此本文改进了一种基于 GRU 和 PCNN 的模型, 该模型可以有效解决传统模型的不足, 通过结合 GRU 模型和 PCNN 模型的优点, 实验结果表明该方法相比传统方法效果极佳, 可以有效实现对电力系统知识抽取.

关键词: 知识抽取; 神经网络; 电力系统; PCNN 模型

引用格式: 宋厚岩,王汉军.基于 GRU 和 PCNN 的电力知识抽取.计算机系统应用,2021,30(9):200–205. http://www.c-s-a.org.cn/1003-3254/8046.html

Knowledge Extraction in Electric Power Based on GRU and PCNN

SONG Hou-Yan[1,2], WANG Han-Jun[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: The main part of drawing the knowledge map of electrical power systems is the extraction of power knowledge. In the traditional supervised-learning-based single neural network models, CNN performs well in extracting the most important local features but is not suitable for processing sequence input, and RNN is strong in tackling serialization tasks but weak in extracting important features. To solve these problems, this study puts forward a model based on GRU and PCNN. Compared with traditional models, this model combining the advantages of the GRU helped model and the PCNN model can obtain impressive results and effectively extract the knowledge of electrical power systems.

Key words: knowledge extraction; neural network; electrical power system; PCNN model

1　引言

随着大数据时代的到来, 电力系统每天产生大量的数据, 但是这些海量的数据却尚未得到合理的挖掘, 导致很多潜在的价值没有被开发, 如何管理和挖掘这些数据的价值成为了电力系统改革和发展的新动力. 为了更好地挖掘电力系统的海量数据, 电力系统与人工智能的结合呼之欲出. 电力知识的提取对于构建电力系统知识图谱来说十分重要. 简单来说, 就是从电力系统产生的海量数据中提取电力知识三元组, 并利用图数据库技术构建电力系统知识图谱, 以方便海量电力数据的管理和挖掘.

知识图谱 (knowledge graph) 是由谷歌在 2012 年正式提出的概念, 重点在于提高搜索引擎的智能化和效率. 知识图谱实际上是一个语义网络, 语义网络就是用节点表示实体或属性, 边表示实体之间、实体与属性之间的各种语义关系[1−3]. 其中, 实体是指客观存在

① 收稿时间: 2020-11-27; 修改时间: 2020-12-28; 采用时间: 2021-01-07; csa 在线出版时间: 2021-09-02

于现实世界中且可区分的物体或事物; 属性是描述实体特征的信息, 如面积和长度等. 关系是知识图谱最重要的特征, 据此才能实现万事万物的互联, 从而支持各种应用, 如语义理解和信息检索等.

所谓知识抽取, 就是提取来自不同来源、不同结构的数据, 形成知识最终存到知识图谱的过程. 知识抽取的大致任务如图 1 所示.

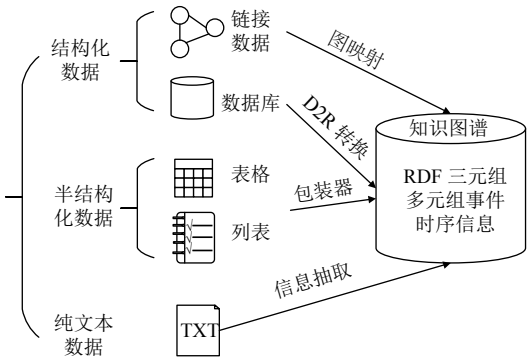

图 1 知识抽取分类图

传统的知识抽取方法包括基于统计机器学习的方法、基于自然语言处理的方法和基于深度学习的方法. 对于基于深度学习的知识抽取方法, 专家们在 2014 年逐渐提出将 CNN 用于知识抽取, 提出了 CNN、RNN 和 LSTM 等一系列方法模型[4–6]. 深度学习网络模型知识提取方法的一般过程是在分词后输入词向量信息, 然后通过卷积神经网络自动学习, 这大大减少了人工标注语料库的人力物力, 同时避免了传统方法中使用自然语言处理等方法造成的累积误差的传播语等问题.

然而目前大多数的知识抽取方法中都使用单一的 CNN 和 RNN 模型, 然而 CNN 擅长的是对于知识的实体局部最重要特征的抽取, 但是却不是十分适合对于序列输入的处理; 反之, RNN 相比于 CNN 而言, 在任意长度的序列化处理上有更好的优势, 但是提取实体局部重要特征却显得不够充分, 所以针对于以上两种模型的不足, 本文改进一种基于 GRU 和 PCNN 模型的电力知识抽取方法. 通过将 GRU 模型与 PCNN 模型的结合, 可以将二者的优点更完美的结合起来, 最终通过实验验证, 该模型对于电力知识抽取相比于单一的 CNN 模型、RNN 模型等神经网络模型有更好的效果[7].

2 电力知识抽取模型

本文所采用的基于深度学习的 GRU 神经元和 PCNN 模型电力知识抽取模型结构图如图 2 所示.

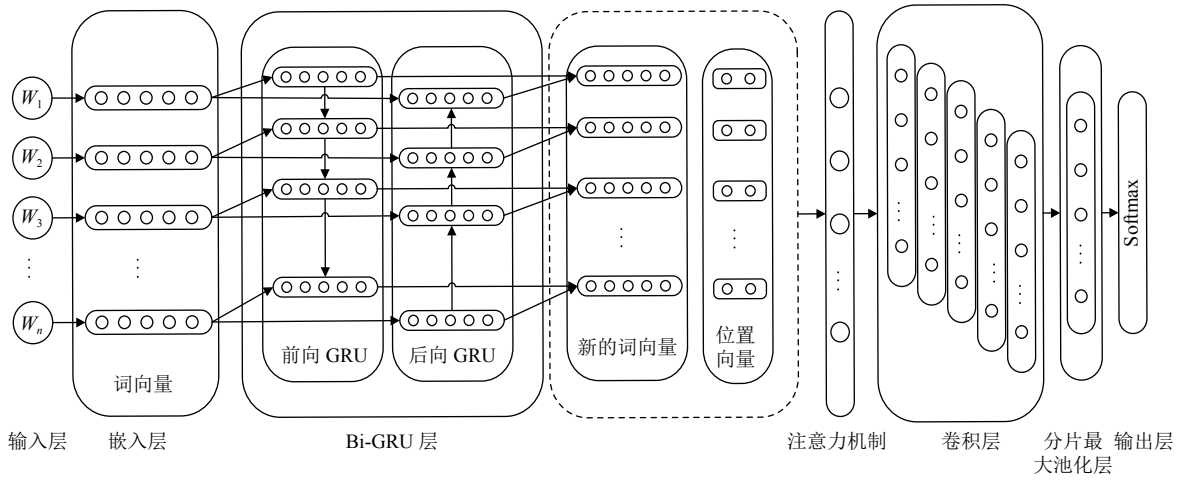

图 2 电力知识抽取模型结构图

本文所设计的电力知识抽取模型的训练主要包括以下几部分:

(1) 输入层: 输入层的主要任务是将电力数据输入到电力知识抽取模型中;

(2) 嵌入层: 通过分词工具和 Word2Vec 词向量工具对输入层输入的数据处理最终得到词向量 W, 将其作为 GRU 模型的输入层;

(3) GRU 层: 利用 GRU 层计算输入层得到的字向量信息, 得到包含新信息的新的词向量. 新词向量通过词向量与词向量的位置向量相拼接所得到的. 新词向

量不仅包含词向量本身的语义信息, 还包含词向量的位置信息;

(4) PCNN 层: 通过对上层特征向量进行实体划分, 通过分段最大池化分从划分结果中提取最重要的局部特征信息, 最终得到句子特征向量 P;

(5) 输出层: 将通过 PCNN 层得到的特征向量 P 输入到 Softmax 分类器中得到最终的电力知识抽取结果.

2.1 GRU 神经元

GRU 是一种 RNN 网络, 该模型的优点在于它适用于处理序列数据, 因此被广泛应用于语音处理、自然语言处理等方向. 它的内部结构与 LSTM 网络的内部结构非常相似, 但是 GRU 模型可以有效解决传统 RNN 网络存在的问题例如梯度爆炸和梯度消失等. 并且与 LSTM 网络相比, GRU 的优点是在于去除细胞单元状态, 传输信息实在隐藏状态下进行的. 尽管 GRU 模型与 LSTM 模型相似, 但前者结构更加简单, 参数更少. 相比之下, GRU 模型就更容易训练. GRU 模型只包含两个门结构, 结构简单, GRU 神经元结构如图 3 所示.

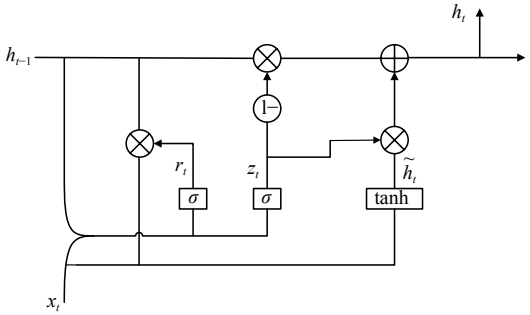

图 3 GRU 神经元结构

如图 3 所示, x_t 为输入数据, h_t 为 GRU 模型的输出, r_t、z_t 分别代表 t 时刻的重置门与更新门, 具体公式如下:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{1}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \times h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \tag{4}$$

$$y_t = \sigma(W_o \cdot h_t) \tag{5}$$

其中, σ 为 Sigmoid 函数, [] 表示两个向量相连, ×表示矩阵的 Hadamard 积, W_r, W_z, W_h 分别为重置门, 更新门, 以及候选隐匿状态的权重矩阵, 其中 $\tilde{h}_t$ 为 t 时刻的候选状态.

在解决自然语言处理时, 通过引用双向循环来处理序列化数据[8–10]. 本文所采用双向 GRU 模型作为电力知识抽取的一部分, 其模型结构分为 3 层, 分别是输入层 (input layer)、隐藏层 (hidden layer)、输出层 (output layer)[11], 其模型如图 4 所示.

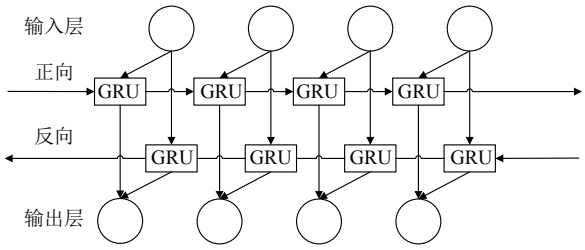

图 4 双向 GRU 模型

在模型的训练过程中, 前向传播过程中公式表示如式 (6)~式 (8) 所示:

$$W_r = W_{rx} + W_{rh} \tag{6}$$

$$W_z = W_{zx} + W_{zh} \tag{7}$$

$$W_{\tilde{h}} = W_{\tilde{h}x} + W_{\tilde{h}h} \tag{8}$$

在输出层部分, 输出层的输入为:

$$y_t^i = W_o h \tag{9}$$

输出层的输出为:

$$y_t^o = \sigma(y_t^i) \tag{10}$$

在得到最终的输出后, 就可以计算出整个网络传递的损失, 单个样本某时刻的损失为:

$$E_t = \frac{1}{2}(y_d - y_t^o)^2 \tag{11}$$

从而可得到单个样本的在所有时刻的损失为:

$$E = \sum_{t=1}^{T} E_t \tag{12}$$

采用后向误差传播算法来学习网络, 先要求出损失函数对各参数的偏导:

$$\frac{\partial E}{\partial W_o} = \delta_{y,t} h_t \tag{13}$$

$$\frac{\partial E}{\partial W_{zx}} = \delta_{z,t} x_t \tag{14}$$

$$\frac{\partial E}{\partial W_{zh}} = \delta_{z,t} h_{t-1} \tag{15}$$

$$\frac{\partial E}{\partial W_{\tilde{h}x}} = \delta_{zt} x_t \tag{16}$$

$$\frac{\partial E}{\partial W_{\tilde{h}h}} = \delta_t (r_t \cdot h_{t-1}) \tag{17}$$

$$\frac{\partial E}{\partial W_{rx}} = \delta_{r,t} x_t \tag{18}$$

$$\frac{\partial E}{\partial W} = \delta_{r,t} h_{t-1} \tag{19}$$

其中, 各中间参数为:

$$\delta_{y,t} = (y_d - y_t^o) \cdot \sigma' \tag{20}$$

$$\begin{aligned}\delta_{h,t} = {} & \delta_{y,t} W_o + \delta_{z,t+1} W_{zh} + \delta_{t+1} W_{\tilde{h}h} \cdot r_{t+1} \\ & + \delta_{h,t+1} \cdot W_{rh} + \delta_{h,t+1} \cdot (1 - z_{t+1})\end{aligned} \tag{21}$$

$$\delta_{z,t} = \delta_{t,h} \cdot (\tilde{h}_t - h_{t-1}) \cdot \sigma' \tag{22}$$

$$\delta_t = \delta_{h,t} \cdot z_t \cdot \phi' \tag{23}$$

$$\delta_{r,t} = h_{t-1} \cdot [(\delta_{h,t} \cdot z_t \cdot \phi') W_{\tilde{h}h}] \cdot \sigma' \tag{24}$$

从训练结果来看, 通过以上公式可知神经网络的结果最终趋于收敛, 因此本文选择 GRU 来用于电力知识的抽取.

2.2 PCNN 模型

PCNN (Piece-wise-CNN) 模型是 CNN 模型的一种, CNN 模型全称是卷积神经网络模型, 近年来被应用于自然语言处理例如情感分析, 文本分类等任务, 而 PCNN 是近年来被发现用于关系抽取的十分经典的一个模型, 也是目前公认的效果较好的抽取模型, 因此本文选用 PCNN 进行电力知识提取是可行的[12–14].

PCNN 模型主要由以下几部分组成, 其结构如图 5 所示.

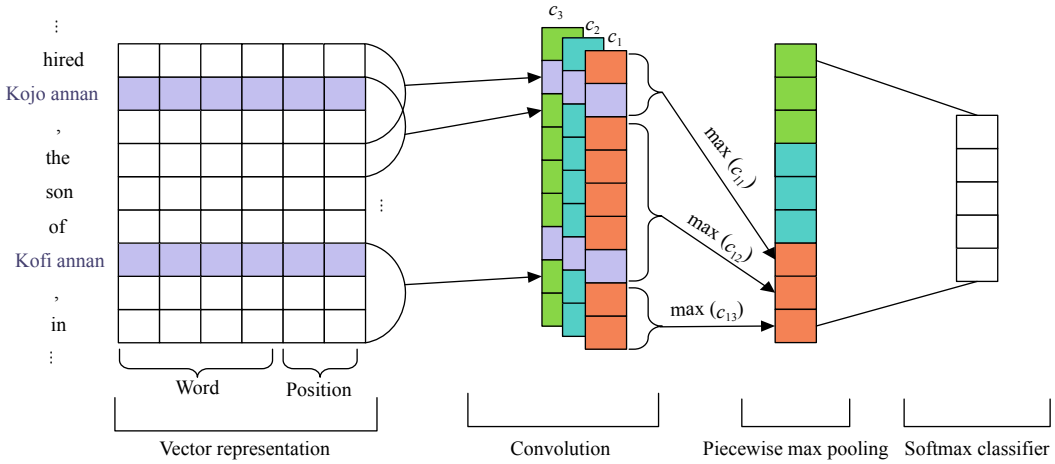

图 5 PCNN 模型

(1) 向量表达: 句子的向量表达为两部分拼接结果: 词嵌入和位置嵌入. 使用 Skip-gram 方法预训练词向量; 使用 Position embedding 表示句子单词到两个实体的相对距离. 随机初始化两个位置嵌入矩阵, 句向量表示为:

$$S = \mathbb{R}^{s \times d} \tag{25}$$

其中, s 为句子长度 (单词数), $d=d_w+d_p \times 2$.

(2) 卷积层: 对于长度为 s 的句子, 首尾填充 $w-1$ 长度, 则卷积核 w 的输出为:

$$c \in \mathbb{R}^{s+w-1},\ c_j = wq_{j-w+1:j},\ 1 \le j \le s+w-1 \tag{26}$$

若使用 n 个卷积核, 则卷积操作的输出为:

$$C = \{c_1, \cdots, c_n\},\ c_{ij} = w_i q_{j-w+1:j},\ 1 \le i \le n \tag{27}$$

(3) 分段最大化池化层: 卷积层输出维度为 $R^{n\times(s+w-1)}$, 输出维度依赖于句子的长度, 为了便于下游任务, 卷积层的输出必须独立于序列长度, 一般采用池化操作, 使用单一最大池化无法捕获两个实体的结构信息特征, PCNN 使用分段最大池化代替单一最大池化, 分段最大池化输出长度为 3 的向量:

$$p_i = \{p_{i1}, p_{i2}, p_{i3}\},\ p_{ij} = \max(c_{ij}),\ 1 \le i \le n, 1 \le j \le 3 \tag{28}$$

拼接所有卷积核分段池化层输出为 $p_{1:n}$, 静非线性函数输出为 (维度与句子长度无关)

$$g = \tanh(p_{1:n}), g \in \mathbb{R}^{3n} \tag{29}$$

(4) Softmax 层: 首先将输出转化为类别分数 (Softmax 转换为类别概率)

$$o = W_1 g + b, W_1 \in \mathbb{R}^{n1\times 3n}, o \in \mathbb{R}^{n1} \tag{30}$$

(5) 多实例学习: 为降低数据标注错误的影响, PCNN 使用多实例 (半监督) 学习. 考虑包含 T 个包的训练集$\{M_1, M_2, \cdots, M_T\}$, 其中 $M_i=\{m_i^1, m_i^2, \cdots, m_i^{q_i}\}$, 包中 q_i 个不同实体互为独立. 对于实例m_i^j, 神经网络 Θ 输出向量 o, 其中第 r 个关系对应的概率为:

$$p(r|m_i^j;\theta) = \frac{e^{o_r}}{\sum_{k=1}^{n_1} e^{o^k}} \tag{31}$$

将目标函数定义为极小化每个包的损失, 从而降低包中数据标注错误的影响. 每个包的标签已知, 包中实例标签未知, 训练过程中将包中实例在包标签上的最大概率作为预测输出, 则目标函数定义为:

$$J(\theta) = \sum_{i=1}^{T} \log p(y_i|m_i^j;\theta) \tag{32}$$

$$j^* = \arg\max_j p\left(y_i|m_i^j;\theta\right),\ 1 \le j \le q_i \tag{33}$$

整个过程如算法 1 所示.

算法 1. 多实例学习

1. 初始化 θ. 将包按照 b_s 的大小划分成小批量.
2. 随机选择一个小批量, 然后将包逐一送进网络.
3. 根据式 (33) 在每个包中寻找第 j 个实例 $m_i^j(1\le i\le b_s)$.
4. 通过 Adadelta 算法, 基于$m_i^j(1\le i\le b_s)$的梯度更新参数 θ.
5. 重复步骤 2~步骤 4, 直至收敛或训练轮数达到最大轮数.

3 实验分析

3.1 实验数据及预处理

本文数据集来源于某电网电力系统的电力调度产生的数据. 该数据集包括训练集和测试集两个部分. 其中训练集包括 32.6 k 个句子, 1478.6 k 个字; 测试集包括 2.6 k 个句子, 98.5 k 个字. 对于电力关系的分类, 通过从电力领域专业词典中获取实体以及通过百度百科爬取相关电力词汇及信息, 通过去除重复信息, 选取出现频率最高并且符合实际电力领域的实体关系如表 1 所示.

3.2 评价指标

电力知识抽取的评价指标有准确率 (P)、召回率 (R) 和 $F1$ 值. 其中计算方式为:

$$P = \frac{T_p}{T_p + F_p} \times 100\% \tag{34}$$

$$R = \frac{T_p}{T_p + F_n} \times 100\% \tag{35}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{36}$$

其中, 参数分别定义为: T_p 为模型识别正确的实体个数, F_p 为模型识别出的不相关实体的个数, F_n 为相关的实体但是模型没有识别出个数.

表 1 关系分类

序号	关系分类
1	属性
2	父类
3	子类
4	属于
5	区别
6	功能
7	发明

3.3 实验设置

为了验证模型的效果, 进行如下实验设置, 首先应用 Word2Vec 词向量工具进行训练, 模型选用 Skip-gram 模型. 词向量训练参数如表 2 所示.

表 2 Word2Vec 训练参数

参数	含义	参数值
Size	属性	50
Window	窗口大小	5
Sample	随机采样	0
Softmax	分类	1
Min_count	最少词频	5
Cbow	Cbow模型	0

分别在 CNN、RNN、GRU、PCNN 以及本文提出的 GRU+PCNN 模型进行实验. 为了保证结果的公平性, 所有模型均采用同样的训练集数据, 并且所有模型的参数保持一致. 训练方式为训练 CNN、RNN、GRU、PCNN 及 GRU+PCNN 模型的所有参数, 参数设置如表 3 所示, 在训练的过程中记录下每一轮的训练后的准确率.

3.4 实验结果分析

通过上述实验, 5 种不同模型在相同的测试集下得到的准确率、召回率、$F1$ 值如表 4 所示.

从本实验结果中, 可以发现, 本文所采用的基于 GRU 和 PCNN 模型无论在准确率、召回率及 $F1$ 值都比其他单一神经网络模型结果更优. 由于 PCNN 模型通过分片最大化池及多实例学习, 可以更好的获取上下文信息, 使得 PCNN 比 GRU 可以发挥出发更好的优

势; 另一方面, GRU 模型具有良好的学习依赖关系的能力. 综上, 本文所才用的 GRU+PCNN 模型比其他模型有明显的提升的原因就在于 GRU+PCNN 模型融合了两个单一模型的优点, 不仅具备了 GRU 的优点, 例如学习长依赖关系的能力, 还具有 PCNN 模型的优点如提取局部重要特征的能力, 因此使得本文所选用的 GRU+PCNN 模型比其他单一模型无论是准确率还是召回率都有明显的提升.

表 3 GRU+PCNN 模型训练参数

参数	含义	参数值
Size	词向量维度	100
Window	窗口大小	3
Max_length	句子最大长度	80
Filter_num	卷积核数	250
Epoch	训练迭代次数	150
Batch	每次训练样本数	50
Dropout_rate	丢失率	0.5
Learning_rate	学习率	0.001

表 4 各模型对比结果 (%)

模型	准确率	召回率	*F*1值
CNN	83.25	56.90	67.59
RNN	82.69	57.86	68.08
GRU	86.55	61.03	71.58
PCNN	84.76	60.55	70.64
GRU+PCNN	89.90	61.25	72.86

4 总结与展望

本文通过将深度学习模型 PCNN 和双向 GRU 的组合模型对于电力领域进行知识抽取. 该模型相比于传统的基于监督学习、机器学习及单一神经网络模型等存在的问题加以改善, 使得该模型既具有 PCNN 模型的优点, 例如 PCNN 模型通过分段最大池化及多实例学习来提取句子局部最重要特征的优点等, 与此同时还具备 GRU 模型的优点, 例如 GRU 可以学习长序列依赖关系的能力. 最终训练得到的实验结果相比于传统模型及单一神经网络模型都有较好的结果, 未来的工作将尝试进一步优化模型来提高电力知识抽取的效果.

参考文献

1 Ruiz LGB, Pegalajar MC, Arcucci R, *et al*. A time-series clustering methodology for knowledge extraction in energy consumption data. Expert Systems with Applications, 2020, 160: 113731. [doi: 10.1016/j.eswa.2020.113731]

2 马忠贵, 倪润宇, 余开航. 知识图谱的最新进展、关键技术和挑战. 工程科学学报, 2020, 42(10): 1254–1266.

3 李涛, 王次臣, 李华康. 知识图谱的发展与构建. 南京理工大学学报, 2017, 41(1): 22–34.

4 王仁武, 袁毅, 袁旭萍. 基于深度学习与图数据库构建中文商业知识图谱的探索研究. 图书与情报, 2016, (1): 110–117.

5 Ivanisenko TV, Saik OV, Demenkov PS, *et al*. ANDDigest: A new web-based module of ANDSystem for the search of knowledge in the scientific literature. BMC Bioinformatics, 2020, 21(S11): 228. [doi: 10.1186/s12859-020-03557-8]

6 俞琰, 陈磊, 姜金德, 等. 融合论文关键词知识的专利术语抽取方法. 图书情报工作, 2020, 64(14): 104–111.

7 王明波, 王峥, 邱秀连. 基于双向 GRU 和 PCNN 的人物关系抽取. 电子设计工程, 2020, 28(10): 160–165.

8 吴超, 王汉军. 基于 GRU 的电力调度领域命名实体识别方法. 计算机系统应用, 2020, 29(8): 185–191. [doi: 10.15888/j.cnki.csa.007595]

9 Gu XD. Feature extraction using unit-linking pulse coupled neural network and its applications. Neural Processing Letters, 2008, 27(1): 25–41. [doi: 10.1007/s11063-007-9057-6]

10 龚乐君, 刘晓林, 高志宏, 等. 基于双向 GRU 和 CNN 的药物相互作用关系抽取. 陕西师范大学学报 (自然科学版), 2020, 48(6): 108–113.

11 王易东, 刘培顺, 王彬. 基于深度学习的系统日志异常检测研究. 网络与信息安全学报, 2019, 5(5): 105–118.

12 秦磊. 新闻文本中人物关系抽取的研究 [硕士学位论文]. 武汉: 武汉邮电科学研究院, 2020.

13 Liu Q, Zhao XL, Yang HP, *et al*. Image feature extraction and retrieval of the Euler number to Chinese herbal medicine based on PCNN. Proceedings of the 3rd International Conference on Computer Graphics and Digital Image Processing (CGDIP 2019). Rome, Italy. 2019. 8.

14 刘伟, 陈鸿昶, 黄瑞阳. 基于 Tree-based CNN 的关系抽取. 中文信息学报, 2018, 32(11): 34–40. [doi: 10.3969/j.issn.1003-0077.2018.11.005]