

基于事理知识图谱的舆情推演方法^①



于 强¹, 徐志栋², 时 斌³, 魏 伟⁴, 任鹏程¹

¹(中国石油大学(华东) 计算机科学与技术学院, 青岛 266580)

²(中国人民解放军国防大学 国家安全学院, 北京 100091)

³(青岛海尔空调电子有限公司, 青岛 266101)

⁴(青岛海尔智能技术研发有限公司, 青岛 266101)

通讯作者: 于 强, E-mail: yuqiangworkmail@163.com

摘 要: 一直以来舆情态势发展的多元性、复杂性使其难以有效管控, 一些负面舆情会激化矛盾, 给社会安定带来不利影响. 提出了一种基于事理知识图谱的舆情事件推演方法, 通过神经网络挖掘事件因果逻辑, 连接因果事件构成事理知识图谱. 向量化事件节点以融合归并相似节点降低图谱冗余, 增强图谱泛化性. 根据事理知识图谱反映的发展逻辑对目标舆情事件的演化趋势进行预测. 以自然灾害舆情事件为例, 实验结果表明提出的方法能够有效预测舆情事件发展方向, 可以为舆情监管提供一定支持.

关键词: 事理知识图谱; 舆情推演; 因果逻辑; 舆情监管

引用格式: 于强, 徐志栋, 时斌, 魏伟, 任鹏程. 基于事理知识图谱的舆情推演方法. 计算机系统应用, 2021, 30(4): 25-31. <http://www.c-s-a.org.cn/1003-3254/7892.html>

Public Opinion Deduction Based on Event Logic Graph

YU Qiang¹, XU Zhi-Dong², SHI Bin³, WEI Wei⁴, REN Peng-Cheng¹

¹(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

²(Institute of National Security, China People's Liberation Army National Defence University, Beijing 100091, China)

³(Qingdao Haier Air Conditioning Electronics Co. Ltd., Qingdao 266101, China)

⁴(Qingdao Haier Intelligent Technology R&D Co. Ltd., Qingdao 266101, China)

Abstract: The diverse and complex trend of public opinion has long made it difficult to manage. Negative public opinion will intensify contradictions, bringing adverse effects to social stability. Then a method of public opinion deduction based on the event knowledge graph is proposed. The causal logic of the event is mined through the neural network, and the event knowledge graph is drawn after causal events are connected. Vectorized event nodes can merge into similar nodes to reduce map redundancy while enhancing map generalization. Besides, the evolution of target public opinion events can be predicted based on the deductive logic indicated in the event knowledge graph. With a public opinion event related to a natural disaster as an example, the experimental results prove that the proposed method can reliably predict the trend of the event, supporting public opinion supervision.

Key words: event logic graph; public opinion deduction; causal logic; public opinion supervision

1 引言

信息技术的发展促进交流方式的转变, 众多网络媒体、社交平台成为大众了解信息、获取信息的重要

来源, 催生了网络舆情这一社会舆论独特表现形式的产生与发展. 网络舆论具有强大的社会监督能力^[1,2], 但如果网络舆论失控, 将会给社会安定带来不利影响. 在

^① 基金项目: 山东省自然科学基金 (ZR2019MF049)

Foundation item: Natural Science Foundation of Shandong Province (ZR2019MF049)

收稿时间: 2020-08-13; 修改时间: 2020-09-29; 采用时间: 2020-10-13; csa 在线出版时间: 2021-03-30

众多类型的人类知识中,事理逻辑是一种非常重要且普遍存在的知识,许多人工智能应用依赖于对事理逻辑知识的深刻理解,但目前的研究缺少针对舆情事件因果动态演化过程的分析,难以对舆情事件发展方向进行有效预测^[3].本文依据采集的舆情数据挖掘因果事件逻辑,构建事理知识图谱,通过文本向量化融合增强事理知识图谱的泛化性.针对目标事件,实现了根据事理知识图谱中相似事件的演化方向,预测其未来发展.

2 相关工作

目前在网络舆情事件推演方面已经出现过诸多研究,前期学者们多利用模糊推理作为演化规则来探究舆情的演化规律.比如张春娇^[4],党小超等^[5]分别考虑信息在传递过程中普遍存在模糊性的特点,结合元胞自动机理论和模糊推理算法建立了网络舆情传播的模糊元胞自动机模型;Ding等^[6]利用模糊元胞自动机分析了不同观点持有者对舆情发展的影响.然而基于推理规则的方法往往停留在对舆情热度、情感等表象的研究,忽视了核心舆情事件发展规律,泛化性难以保证.近年来得益于计算机技术有力发展,学者们开始运用大数据、人工智能技术研究网络舆情演化规律.比如兰月新等^[7]定性的分析了大数据环境下网民情绪特征和分类,构建了网民情绪演化机理微分方程模型分析网民情绪演化趋势;曾子明等^[8]等构建了基于BP神经网络的舆情热度趋势预测模型用于预测突发传染病事件的发展趋势;Yang等^[9]利用多类别支持向量机进行观点挖掘以及情感分析,实现了对舆情的趋势以及热度预测,但该类方法在可解释性上存在欠缺.

哈尔滨工业大学刘挺教授团队率先提出“事理图谱(Event Logic Graph, ELG)”^[10,11]概念,其本质是事件逻辑知识库,用于揭示现实世界事件的演化模式和发展逻辑,对于认识人类行为和社会发展变化规律具有重要的意义.目前基于事理图谱进行舆情事件预测研究正处于起步阶段,单晓红等^[12]、夏立新等^[13]、Li等^[14]在这一领域做出了一些探索,但在舆情逻辑事件抽取与泛化方面仍有待加强.本文在传统通过模式匹配抽取事件基础上,研究了基于神经网络的事件识别与抽取方法,优化了事理知识图谱中边权重计算方式,实验结果证明本文提出的舆情推演方法有效,可以较好地揭示舆情事件演化规律,从而为舆情管控提供支持.

3 舆情推演方法

基于事理知识图谱的舆情推演方法如图1所示.

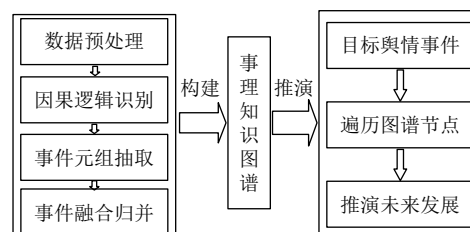


图1 舆情推演流程

首先处理原始舆情语料,识别、抽取出因果事件元组;其次对事件进行融合减少冗余,完成事理知识图谱构建与泛化;最终根据图谱中事件节点的演化规律对目标舆情事件的可能发展动向进行研判.

3.1 因果逻辑事件识别与抽取

本文以因果关系逻辑为基础构建事理知识图谱,将其分为了两个过程.首先对文本进行分析,判断识别是否含有因果逻辑,然后再抽取事件元组.

1) 因果逻辑识别

我们将事件因果关系逻辑识别作为文本分类任务处理,设计了基于BERT的因果逻辑事件识别模型.BERT^[15]是谷歌团队于2018年底发布的基于双向Transformer^[16]的大规模预训练语言模型,在多项自然语言处理任务中获取了最好效果.

我们对标准的BERT模型进行了改进,在BERT模型输出层取得所有输入字符对应的输出向量后对接文本分类器,分类器选择包括长短时记忆网络BiLSTM、循环卷积神经网络RNN,用于对BERT输出的向量再次进行计算,判断其是否含有因果逻辑语义.进一步,我们使用了原始的BiLSTM、RNN以及Transformer模型处理相同的实验数据,以对比分析BERT模型的加入以及不同BERT模型改进方式对结果造成的影响,各个模型的准确率在实验部分给出.实验结果显示BERT-BiLSTM模型能够得到最好的识别分类效果,后续处理分析将基于BERT-BiLSTM模型处理结果进行.

用于因果逻辑识别的BERT-BiLSTM模型如图2所示.

对于任意输入文本序列,在完成数据清洗之后处理为单个字符的形式输入模型,便可自动判断其是否属于因果逻辑性描述.

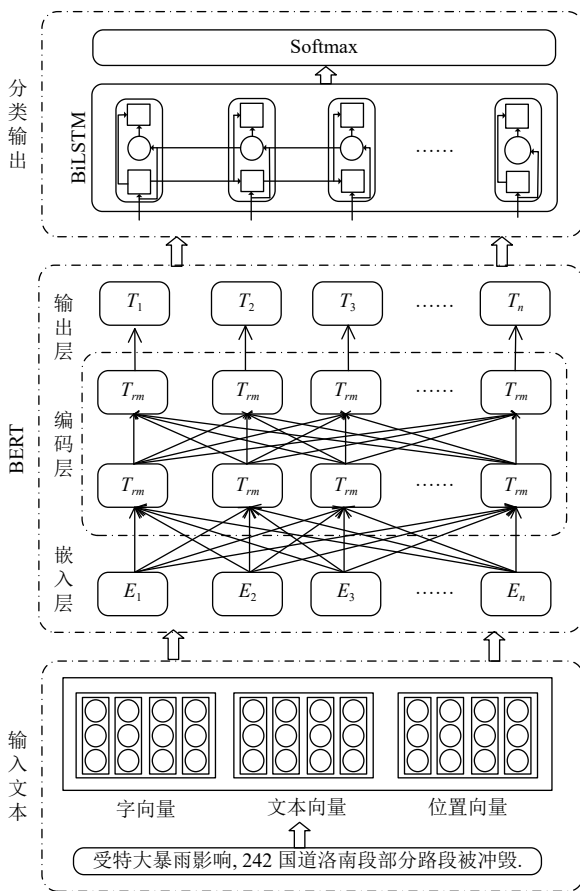


图2 因果逻辑识别

从图2可以明显看到 BERT 模型由嵌入层、编码层、输出层 3 部分构成, 关键部分是双向 Transformer 结构, 实质是一个基于“自注意力机制”的深度学习, 即通过计算同一个句子中的词与词之间的关联程度调整权重系数矩阵以表征词:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

其中, \$Q, K, V\$ 是字向量矩阵, \$d_k\$ 是 Embedding 的维度, 多头注意力机制通过多个不同的线性变化对 \$Q, K, V\$ 进行投影, 通过公式 (2)(3) 将不同 Attention 结果拼接起来.

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_n)W^o \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

表1 标注示例

词语	受	特大	暴雨	影响	,	242	国道	洛南	段	部分	路段	被	冲毁	.
标签	O	B_C_1	E_C_1	O	O	O	O	O	O	B_R_1	I_R_1	I_R_1	E_R_1	O

其中, \$W\$ 是权重矩阵, 由此模型可以实现对文本重点特征的聚焦提取. 编码器结构如图3所示.

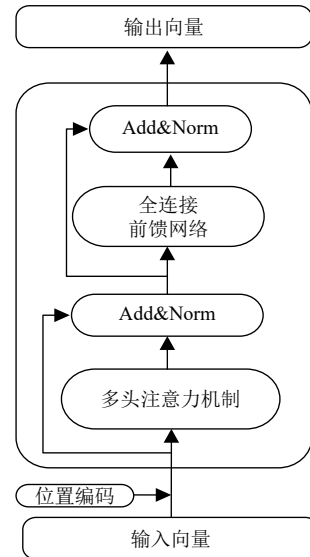


图3 Transformer 编码器

2) 因果逻辑抽取

在筛选得到含有因果逻辑事件描述的文本之后, 本文通过 BiLSTM-CRF^[17] 算法获取因果事件元组. BiLSTM-CRF 算法已被成功应用于实体命名识别工作中, 取得了良好的效果. 而元事件抽取与实体命名识别有许多共通之处, 所以本文将 BiLSTM-CRF 算法引用到元事件抽取过程中. 类比于命名实体抽取方法, 本文采用序列标注任务中经典的 BIO 标注体系^[18] 对数据进行标注, 具体使用的标注标签如下:

- (1) 词语的位置: B (开始), I (内部), E (结束);
- (2) 语义角色信息: C (原因), R (结果);
- (3) 事件的序号: 1-N (每个对应序号为同一事件的因、果);
- (4) 其他词语: O.

例如, 对于“受特大暴雨影响, 242 国道洛南段部分路段被冲毁.”, 标注结果如表1所示. 因果逻辑事件抽取模型需要对输入序列中的每一个词语进行类别判断, 然后为其输出一个类别标签, 标签代表了序列的类别和边界, 元事件抽取过程如图4所示.

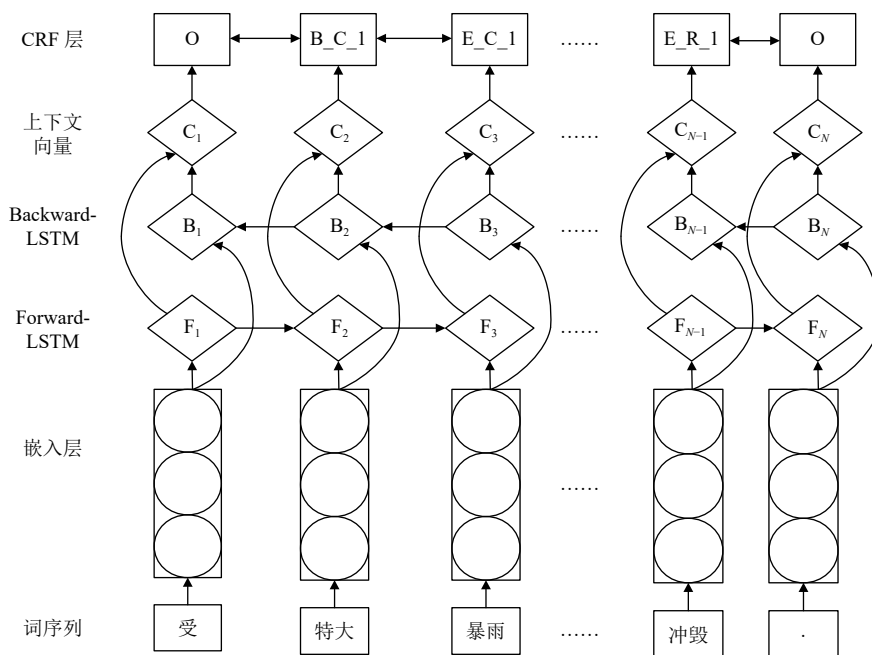


图4 元事件抽取过程

3.2 事理知识图谱构建与泛化

本文将提取到的元事件转化为图谱“因→果”形式,即以事件为节点,因果关系为边构建事理知识图谱.事理知识图谱可以表示为 $EventGraph = \{Nodes, Edges, Transforms\}$, 其中 $Nodes = \{n_1, n_2, \dots, n_k\}$ 为节点, 即元事件集合; $Edges = \{e_1, e_2, \dots, e_k\}$ 为边, 即因果关系, 每一条边都是由原因事件指向结果事件; $Transforms = \{t_1, t_2, \dots, t_k\}$ 为边的权重, 用于计算某一原因事件造成特定结果事件的可能性.

对于抽取结果中重复的因果事件描述可能造成图谱冗余问题, 本文分两种情形处理:

① 重复描述同一舆情事件存在的因果逻辑.

② 属于不同舆情事件但内容相同的因果逻辑, 例如“暴雨引发山体滑坡”事件, 在“2019年7月上中旬长江中下游洪水”, “四川·8·20·强降雨特大山洪、泥石流灾害”等舆情事件中都存在.

我们将事件文本向量化处理, 通过相似度计算解决以上两种问题. 具体方法如下: 对所有舆情事件进行分词处理获得原始语料数据, 使用 Word2Vec^[19] 模型处理所有原始数据, 得到单词向量, 使用事件文本组成词的向量和平均值作为事件向量, 公式为:

$$n_{i\text{vec}} = \frac{\sum w_{i\text{vec}}}{\text{sum}(w_i)} \quad (4)$$

其中, w_i 是事件 n_i 的组成词汇, $w_{i\text{vec}}$ 为对应单词向量, $n_{i\text{vec}}$ 为事件节点 n_i 的向量.

进一步, 计算事件之间向量余弦相似度^[20], 计算公式为:

$$Sim(n_{i,j}) = \frac{n_{i\text{vec}} \cdot n_{j\text{vec}}}{\|n_{i\text{vec}}\| \times \|n_{j\text{vec}}\|} \quad (5)$$

其中, $Sim(n_{i,j})$ 为事件节点 n_i 与 n_j 的相似度. 若两个事件相似度高于预定阈值:

针对情形①, 删除重复描述, 即同一事件内每种因果逻辑只保留一条记录;

针对情形②, 合并为同一事件节点, 并增加对应边的权重. 如图5所示, 阴影节点表示两个事件高度相似, 权重代表某一事件发生过的次数.

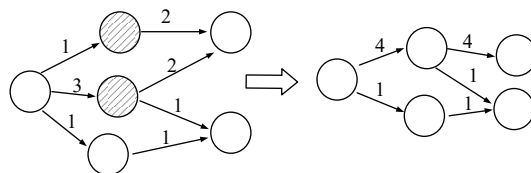


图5 相似事件归并

3.3 舆情事件推演方法

舆情事件推演是在已知某一事件发生之后, 推测它可能导致的后续事件, 本文构建的舆情推演方法具体步骤如图6.

表2 模型结果对比

Model	Accuracy	F1-Score
RNN ^[22]	0.683	0.542
Bi-LSTM ^[23]	0.743	0.709
Transformer	0.792	0.683
BERT-RNN	0.736	0.801
BERT-BiLSTM	0.869	0.768

由表2结果可以看出,基于标准BERT模型改进的BERT-RNN、BERT-BiLSTM相对于基准RNN、Bi-LSTM模型都取得了更好的识别结果,说明在此数据集上BERT模型凭借其创新的训练模式以及参数体量的优势能更加有效的识别出文本特征信息,从而取得更佳的分类型效果。由于本数据集中数据信息都是完整的舆情事件记录,文本长度较大,RNN模型以及BERT-RNN模型都未取得较好的效果,而LSTM由于门控机制的存在,相对RNN能够更加高效的捕捉更长距离的依赖,实现了更好的分类效果,最终BERT-BiLSTM通过结合BERT模型与BiLSTM模型的优势,取得了最优的分类效果。

2) 舆情推演

鉴于在第3.3节中介绍的舆情推演方法与某些推荐算法的工作过程存在异曲同工之处,本文移植了推荐算法的常用评价指标MRR^[24]对舆情推演结果做出评价。

MRR使用正确检索结果值在检索结果中的排名来评估检索系统的性能,是一个国际上通用的对搜索算法进行评价的机制,其计算公式为:

$$MRR = \frac{1}{|Q|} \sum_{o=1}^{|Q|} \frac{1}{rank_i} \quad (6)$$

其中, Q 为样本query集合, $|Q|$ 表示 Q 中query个数, $rank_i$ 表示在第 i 个query中,第1个正确答案的排名。比如某测试集有3个query,结果中的第一个正确答案分别被排在第4,2,5位,则该系统的MRR得分为 $(1/4 + 1/2 + 1/5)/3 = 0.3177$ 。

在本文中基于已构建事理知识图谱为测试事件(因)推测可能后续事件(果),出现多个推测结果情况时则是根据边的权重系数大小进行排序。我们对2020年“南方水灾”数据中的因果信息进行了人工筛选与抽取,共得到166个因果事件对作为测试数据。使用MRR评价指标进行评分,最高准确率得分为0.716,这证明了本文所提出方法的有效性。

同时,本文分析了使用各不同因果识别模型以及不同事件相似度阈值设置下对模型结果造成的影响,图9显示了不同相似度阈值设置下模型推演结果准确率的变化。

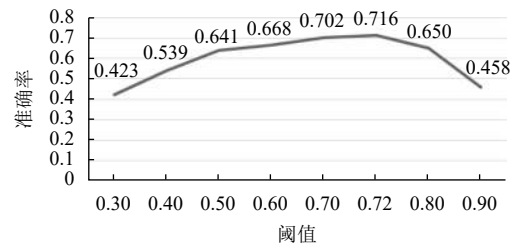


图9 相似度阈值-准确率影响

通过实验结果表明,相似度阈值的变化会对模型推演性能造成一定的影响。若相似度阈值设置过小会造成事件过度匹配;相反,若相似度阈值设置过大会造成事件欠缺匹配。在设置事件相似度计算阈值为0.72时可以在本文数据集上取得最优结果。

本文同时分析了使用不同事件识别模型对最终推演结果的影响,实验过程事件相似度阈值设置为0.72,结果如图10所示。

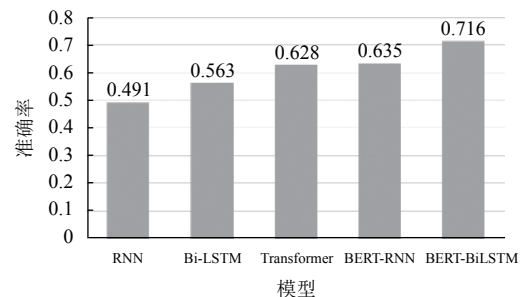


图10 事件识别模型对应模型推演结果准确率

结果表明,不同因果逻辑事件识别模型处理结果的差异进一步影响到了因果元事件抽取效果及事理知识图谱的构建,并最终扩散到模型推演效果。选择更好的因果逻辑事件识别模型可以增强事理知识图谱对于舆情事件逻辑信息的表达能力,从而提高舆情推演结果的准确率。

5 结束语

本文提出了一种基于事理知识图谱的舆情事件推演分析方法,具体介绍了因果逻辑事件识别与提取、事理知识图谱的构建、舆情事件演化分析方法,并通

过实验验证了本文提出方法的有效性 & 先进性。舆情事件分析作为舆情治理的核心问题之一, 研究舆情事件演化过程对于维护社会长治久安具有重要意义。

诚然, 本文工作仍有可以改进之处, 主要在于因果逻辑抽取层面, 未来工作将进一步探讨如何更加准确地对事件边界进行界定。

参考文献

- 1 罗霄峰, 罗万伯, 胡月, 等. 网络舆情治理研究. 通信技术, 2010, 43(4): 81–83. [doi: 10.3969/j.issn.1002-0802.2010.04.028]
- 2 胡蓉. 大数据环境下突发危机事件的网络舆情应对研究. 芜湖日报, 2020-07-17(005).
- 3 张志霞, 郝纹慧, 张二双. 网络舆情驱动下突发事件情景推演研究. 情报科学, 2020, 38(5): 141–147.
- 4 张春娇. 网络舆情传播的模糊元胞自动机模型研究 [硕士学位论文]. 兰州: 西北师范大学, 2014.
- 5 党小超, 张春娇, 郝占军. 基于模糊元胞自动机的网络舆情传播模型研究. 计算机工程, 2014, 40(4): 209–213. [doi: 10.3969/j.issn.1000-3428.2014.04.040]
- 6 Ding CL, Wei CF, Gu TT, *et al.* Study on propagation model of network public opinion based on fuzzy cellular automata. Proceedings of 2013 2nd International Conference on Measurement, Information and Control. Harbin, China. 2014. 1009–1013.
- 7 兰月新, 夏一雪, 刘冰月, 等. 面向舆情大数据的网民情绪演化机理及趋势预测研究. 情报杂志, 2017, 36(11): 134–140. [doi: 10.3969/j.issn.1002-1965.2017.11.021]
- 8 曾子明, 黄城莺. 基于 BP 神经网络的突发传染病舆情热度趋势预测模型研究. 现代情报, 2018, 38(5): 37–44, 52. [doi: 10.3969/j.issn.1008-0821.2018.05.006]
- 9 Yang HL, Lin QF. Opinion mining for multiple types of emotion-embedded products/services through evolutionary strategy. Expert Systems with Applications, 2018, 99: 44–55. [doi: 10.1016/j.eswa.2018.01.022]
- 10 项威. 事件知识图谱构建技术与应用综述. 计算机与现代化, 2020, (1): 10–16.
- 11 王毅, 沈喆, 姚毅凡, 等. 领域事件图谱构建方法综述. 数据分析与知识发现, 2020, 4(10): 1–13.
- 12 单晓红, 庞世红, 刘晓燕, 等. 基于事理图谱的网络舆情事件预测方法研究. 情报理论与实践, 2020, 43(10): 165–170, 156.
- 13 夏立新, 陈健瑶, 余华娟. 基于事理图谱的多维特征网络舆情事件可视化摘要生成研究. 情报理论与实践, 2020, 43(10): 157–164.
- 14 Li ZY, Ding X, Liu T. Constructing narrative event evolutionary graph for script event prediction. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden. 2018. 4201–4207.
- 15 Devlin J, Chang MW, Lee K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805, 2018.
- 16 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing. Red Hook, NY, USA. 2017. 6000–6010.
- 17 Sun L, Rao Y, Lu Y, *et al.* A method of Chinese named entity recognition based on CNN-BILSTM-CRF model. Proceedings of the 4th International Conference of Pioneering Computer Scientists, Engineers and Educators on Data Science. Singapore. 2018. 161–175.
- 18 高冰涛, 张阳, 刘斌. BioTrHMM: 基于迁移学习的生物医学命名实体识别算法. 计算机应用研究, 2019, 36(1): 45–48.
- 19 Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.
- 20 鲜翠琼, 秦学, 朱道恒, 等. 一种图文组合相似度算法的设计与优化. 软件工程, 2020, 23(8): 9–12, 4.
- 21 应急管理部救灾和物资保障司. 应急管理部公布 2019 年全国十大自然灾害. 中国减灾, 2020, (3): 12–15.
- 22 Mikolov T, Kombrink S, Burget L, *et al.* Extensions of recurrent neural network language model. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic. 2011. 5528–5531.
- 23 Wang J, Zhu L, Dai T, *et al.* Deep memory network with Bi-LSTM for personalized context-aware citation recommendation. Neurocomputing, 2020, 410: 103–113. [doi: 10.1016/j.neucom.2020.05.047]
- 24 任函. 基于推理现象识别的答案抽取. 湖北科技学院学报, 2017, 37(4): 132–135. [doi: 10.3969/j.issn.1006-5342.2017.04.030]