

火灾应急领域知识图谱问答方法研究

潘 茹¹, 查 俊^{1,2}

(1.安徽建筑大学电子与信息工程学院, 安徽 合肥 230601;

2.清华大学合肥公共安全研究院, 安徽 合肥 230601)

✉ 768314326@qq.com; zhajun@ahlkj.com



摘 要:火灾是常见的一种灾害,火灾应急知识的获取对个人安全和社会发展至关重要。文章提出一种面向火灾应急领域的知识图谱问答 Pipeline(流水线模型)改进方法。首先对实体识别、实体链接及路径排序阶段的模型采用 Task specific 的思想进行独立训练,结合特征融合算法对知识三元组进行重新排序;其次在路径排序模型中引入 Beam Search(集束搜索)算法;最后采用迁移学习策略,在通讯领域的知识图谱问答语料场景下训练大参数量模型。实验证明,该方法应用于火灾应急领域语料上的准确率为 89.0%,优于传统 Pipeline 方法,问答效果更好。

关键词:火灾应急;知识图谱问答;信息检索;语义匹配

中图分类号:TP391.1 **文献标识码:**A

Research on Q&A Method of Knowledge Graph in Fire Emergency Domain

PAN Ru¹, ZHA Jun^{1,2}

(1.School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China;

2.Hefei Institute of Public Safety Research, Tsinghua University, Hefei 230601, China)

✉ 768314326@qq.com; zhajun@ahlkj.com

Abstract: Fire is one of the most common disasters, and the acquisition of fire emergency knowledge is crucial for personal safety and social development. This paper proposes a method of improving knowledge graph Q&A (Questioning-answering) Pipeline (a pipeline model) for fire emergency domain. Firstly, the models for entity recognition, entity linking and path ranking are trained independently using the idea of Task specific, and the knowledge triplets is re-ranked by combining feature fusion algorithm. Secondly, Beam Search algorithm is introduced in the path ranking model. Finally, transfer learning strategy is adopted to train the large parametric model under the knowledge graph Q&A corpus of general domain. The experimental results show that the proposed method has an accuracy of 89.0% on the corpus of fire emergency field, which is better than the traditional Pipeline method, and has better Q&A performance.

Keywords: fire emergency; knowledge graph Q&A; information retrieval; semantic matching

0 引言(Introduction)

据国家消防救援局的公开数据统计,2021年,全国消防救援队伍共接报处置各类警情195.6万起,其中共接报火灾74.8

万起,占总警情的38.2%,是社会单位常见、危害较大的灾害^[1]。火灾的应急救援行动往往受限于应急人员火灾知识的积累,然而当前火灾应急的相关知识和应急预案常常以多元

化、非结构化的形式分布在互联网和相关书籍中,基于传统的搜索方式和问答系统对特定场景进行精确搜索的难度较大。因此,研究基于知识图谱的问答方法,为用户提供精确的火灾应急知识具有重要意义^[2]。

目前,主流的知识图谱问答方法研究有基于语义解析的方法和基于信息检索的方法两类^[3-6]。前者主要通过规则或结构将问句查询解析为问句语义逻辑形式,然后在知识库中执行该逻辑形式的表达式,以寻找答案集合,但该方法依赖人工解析语义关系,预测未知关系困难。后者通过链接问句主题实体临近的子图,应用排序算法选择最高位置的实体检索答案,该方法泛化性强,但推理过程强依赖上游任务的输出结果,易将错误传播至下游任务,降低了推理过程的可解释性。因此,本文基于传统的信息检索式的知识图谱问答方法,提出一种基于火灾应急领域知识图谱的问答 Pipeline 改进方法,旨在解决下游推理对上游结果的强依赖、答案推理的效率低下,以及小规模知识图谱模型训练的过拟合等问题。

1 火灾应急知识图谱问答方法(Fire emergency knowledge graph Q&A method)

火灾应急知识图谱问答方法主要分为两个阶段:第一个阶段为火灾应急领域问句理解,对火灾应急领域提问进行火灾领域实体识别、火灾领域实体链接,以获取问句中的实体并链接到火灾应急领域知识图谱中;第二个阶段为答案检索与推理,基于第一阶段获取到的火灾领域实体,进行实体路径生成、排序和答案检索,具体内容如下:先针对上一步骤中获取的火灾领域实体在知识库中的子图,基于特定的规则生成路径,再通过路径排序模型为生成的路径进行排序,同时基于特征融合的方法对前期特征进行整合,避免问句理解阶段的错误传播,最后将路径还原为知识图谱中的三元组,经过知识图谱的查询语句检索答案返回给用户。

火灾应急领域知识图谱问答方法框架图如图 1 所示。以问句“请帮我找到森林火灾的应急预案”为例,首先对问句进行实体识别,得到“应急预案”和“森林火灾”两个实体,“应急预案”的图谱流行度较高,其得分为 1.0,“森林火灾”由命名实体识别模型评判得分为 0.58。在传统 Pipeline 方法中,直接将得分较高的“应急预案”作为结果进行唯一输出,从而导致在实体链接及路径排序阶段输出的三元组中,找不到“森林火灾”实体的相关路径,最终得到一个不准确的答案。但是,本文所提方法将实体识别、实体链接及路径排序阶段的特征分数予以保留,即使“森林火灾”实体在上游任务中的得分较低,但也不会舍弃该实体,而是在答案推理阶段通过特征融合进行重新排序,让“森林火灾”这条路径上的三元组获得较高的得分(0.86),从而输出(“森林火灾”-“应急预案”-A)的结果,为用户提供了更精准的回答。其中,图 1 中的火灾领域实体识别、

火灾领域实体链接、候选路径排序模型使用了迁移学习的策略,在外部的知识图谱问答语料上进行了预训练。

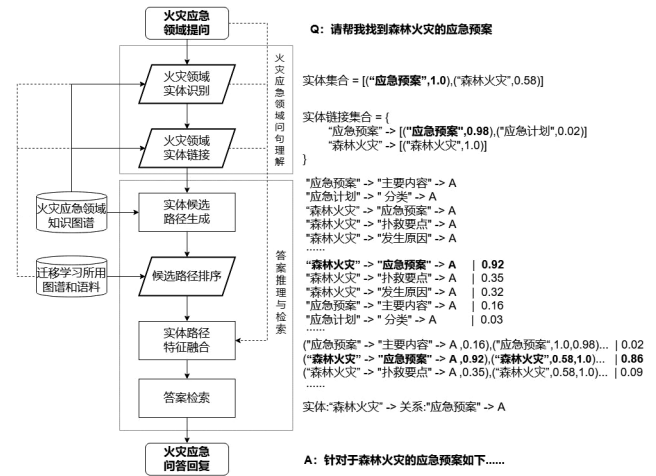


图 1 火灾应急知识图谱问答方法框架图

Fig. 1 Framework of fire emergency knowledge graph Q&A method

2 火灾应急领域问句理解 (Question comprehension for fire emergency domain)

2.1 火灾领域实体识别

本文主要使用 BERT (Bidirectional Encoder Representations from Transformers, 双向 Transformers 编码器)+CRF (Conditional Random Field, 条件随机场)模型、词典匹配及规则提取三种方法联合进行实体的识别抽取。其中, BERT+CRF 模型输入的是将火灾应急领域问句字符串基于 BERT 的词典转换后的字词向量,经过多个 Transformer 编码器计算处理后,输入 CRF 模型中进行序列 BIOES 标注(O 表示非实体;B-M 表示实体词开头的第一个字符或单字符的实体;I-M 表示实体词的中间或结尾字符),经过 CRF 标注处理后输出一个概率矩阵,得到候选实体集合 M_e 及每个实体对应的概率分数。模型结构如图 2 所示。

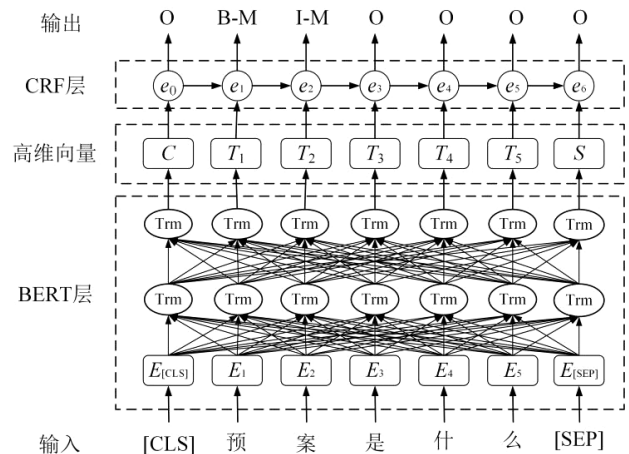


图 2 BERT+CRF 模型结构图

Fig. 2 Structure diagram of BERT+CRF model

词典匹配的方法主要借鉴了最大匹配算法的思想。设 Q 为多个问句的集合, 对于一个给定的问句 $q \in Q$, 由长到短获取问句中的连续子串 str_n , 对句子进行遍历。如果连续子串对应图谱中的某个实体 G_E , 则将该字符串记录到候选实体集合 M_b 中, 并从问句中删除, 以此类推, 直至子串的长度为 0, 则结束对该问句的匹配。词典匹配算法伪代码如下。

输入: 问句集合 Q , 火灾应急领域知识图谱中的实体集合 G

输出: 候选实体集合 M_b

for all $q \in$ 问句集合 Q do

for all $n = q.length$ do

$str_n = q[startidx, startidx + n]$

if $G_E \leftarrow str_n = q[startidx, startidx + n]$ then

$M_b.append(str_n)$

delete(str_n) from q

end if

$n = n - 1$

end for

end for

规则提取法是基于火灾应急领域预设的包括特定地点、人名、专业术语、书名号等信息的特定规则, 提取子串作为实体, 记录到候选实体集合 M_c 中。

最终, 令 i 作为问句的编号, 则 q_i 表示第 i 个问句, 将上述三种方法的实体识别结果进行合并, 得到 q_i 的候选实体集合为 $M_i = M_a \cup M_b \cup M_c$, 多个问句集合 Q 的候选实体集合为 $M = \{M_1, M_2, \dots, M_i\}$, 并记录所有结果。

2.2 火灾领域实体链接

实体链接是将问句中的实体对应到火灾应急知识图谱的实体项中。具体步骤如下: 首先, 进行实体还原, 实现问句中不完整实体的还原。例如, 问句为“药厂火灾该怎么处置?”, 抽取到的实体为“药厂火灾”, 则需要还原为知识图谱中记录的实体项“医药厂火灾”。其次, 基于实体链接词典寻找问句中的实体对应的一个或多个火灾应急知识图谱中的具体实体项。最后, 若问句中的实体链接到多个实体项, 则进行实体消歧处理, 模型结构如图 3 所示, 将问句和链接到的实体分别作为 Text1 和 Text2 输入 BERT 模型中进行处理, 再将 BERT 输出结果的起始位置 [CLS] 标签通过 Concat(连接层)与下游任务进行拼接, 接着经过一个 Dense(全连接层)提取特征, 并添加 Dropout(随机失活层)防止模型训练的过拟合, 通过 Sigmoid 函数完成标签的概率计算, 其核心工作是计算用户提出的问句与实体相关属性的相似度, 计算公式如下:

$$Score(M_i) = SimilarityModel(q, E_i) \quad (1)$$

其中, M_i 是编号为 i 的问句中的候选实体集合, E_i 是对应 M_i 链接到知识图谱后的实体集合, q 是原问句。最终的输出结果

为 $[0, 1]$ 中的某个值, 该值越接近 1, 则说明实体项越符合问句的语境; 反之, 则说明实体项与问句无关, 可能不是用户所提问的实体。

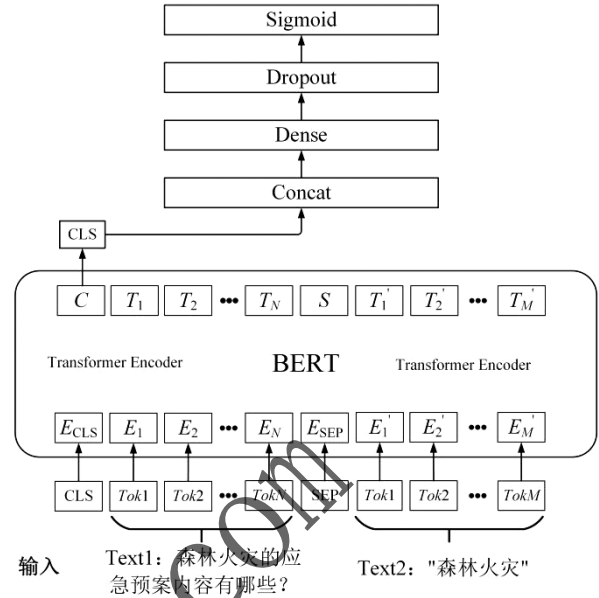


图 3 实体消歧模型结构图

Fig. 3 Structure diagram of entity disambiguation model

3 火灾应急领域答案推理 (Answer reasoning for fire emergency domain)

3.1 候选路径生成

本文采用基于模板生成路径的方式, 定义了 spa 和 apo 两种查询基线, 通过这两种查询基线进行单跳或多跳路径扩展。其中, spa 即实体(主体)-关系/属性-答案, 以实体作为头节点, 经过关系寻找答案; apo 即答案-关系/属性-实体(客体), 以实体作为尾节点, 经过关系寻找答案。本文尽可能地穷举了路径生成的方案, 详见表 1。表 1 中涵盖了常见的单实体单跳路径、单实体多跳路径。

表 1 路径生成模板

Tab. 1 Path generation template

路径集合名	操作	操作对象(集合)
spa	One-hop-s	实体节点
apo	One-hop-o	实体节点
sppa	One-hop-s	spa
spopa	One-hop-o	spa
opspa	One-hop-s	apo
oppa	One-hop-o	apo
spppa	One-hop-s	sppa
apppo	One-hop-o	sppa

3.2 候选路径排序

本文路径排序模型结构与实体链接阶段的实体消歧模型类似, 但模型的输入不同, 这里将问句和候选路径分别作为

Text1 和 Text2。需要说明的是,该模型仅计算实体候选路径与原问句的相似度,并且不关心它所对应的实体是否是问句的中心实体。

考虑到候选路径过多时会导致问答系统的响应效率低下,因此引入 Beam Search 算法针对路径排序模型进行优化。首先重置一个路径集合 P_0 , i 表示当前进行的跳数,当进行第 i 跳时,基于实体路径生成方法,从实体临近的子图中寻找三元组路径并与 P_{i-1} 的路径结合;其次将当前跳生成的每一条路径与原问句计算语义相似度,记录其得分;接着将其得分进行排序,保留 k 个得分排名靠前的路径,将其他得分较低的当前跳路径删除;最后到达最大跳数时,停止循环。Beam Search 优化算法伪代码如下。

输入:问句集合 Q , 候选实体路径集合 P

输出:候选实体路径分数 S_p

for all $q \in Q$ 问句集合 Q do

$P_0 = \emptyset$

for $i \in [1, n_{\max}]$ do

$P_i = \text{findTriples}(P_{i-1}, E_i, G)$

for all $p \in P_i$ do

$S_p = \text{SimilarityModel}(q, p)$

end for

$\text{rankPathByScore}(P_i)$

for all $p \in P_i$ do

if $\text{rankIndex}(S_p) > k$ then

delete(p) from P_i

end for

end for

end for

在本文领域中的具体应用为当目前扩展的关系跳数为 1 跳时,直接执行相似度计算模型,对所有单跳路径进行相似度计算,并记录得分;当目前扩展的关系跳数为 2 跳及以上时,保留前一跳排名前 3 名的路径,在这些路径上进行跳数扩展;约束最大的扩展跳数目前为 3 跳。

3.3 特征融合

在传统的 Pipeline 方法中,对问句进行上游任务处理时就会直接提取核心实体作为唯一确定的问句实体,从而导致问答系统十分依赖实体识别阶段模型的精确度,一旦模型输出的实体出现错误,则会将错误传播至实体链接及实体路径排序阶段,那么问答系统下游任务进行的工作将徒劳无益,最终难以找到正确的答案。

在本文提出的改进 Pipeline 方法中,采用 Task specific(任务专一化)思想,每个任务模块都专注其当前任务,只记录模型当前输出的实体集合及评分结果,不会直接以模型评分的高低确定某一个实体或路径。即本文所提方法将实体识别、实体链接和路径排序中每一个实体及其路径的得分予以保留,并将其作为置信度进行特征融合重新排序。此方法可有效防止实体识别、实体链接阶段出错时,基于错误的实体找到错误的路径和答案的情况。特征融合示意图如图 4 所示。

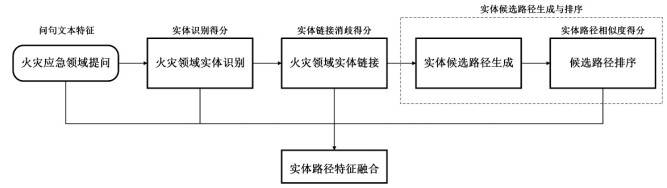


图 4 特征融合示意图

Fig. 4 Schematic diagram of feature fusion

本阶段将当前生成的所有候选路径集合中所对应的相关特征进行融合,其中针对每一条候选路径定义了 6 个相关特征。

(1) S_1 表示火灾领域实体识别阶段的三种方法联合抽取到的实体特征得分。对于通过 BERT+CRF 模型得到的实体,记其特征得分为 BERT+CRF 模型的预测分数;对于通过词典匹配和规则提取方法得到的实体,本文直接对该特征进行赋值,记其特征得分为 1。

(2) S_2 表示实体链接阶段中获取的实体特征得分。若火灾领域实体进行链接时只链接到一个实体项,则认为该火灾应急知识图谱中的实体是唯一的,具体实体项,记该实体的特征得分为 1,否则记为基于实体消歧模型给出的得分。

(3) S_3 属于路径模型特征得分,直接记为候选路径排序模型给出的分数,即计算问句与候选路径的语义相似度。

(4) S_4 属于路径基础特征得分,将候选路径中的字符长度与问句中的字符长度进行比值计算。设问句中的字符长度为 n ,候选路径中的字符长度为 m ,则 $S_4 = m/n$ 。

(5) S_5 属于路径基础特征得分,表示用户问句与候选路径的字符级相似度。采用 Jaccard 相似度计算方法,设 A 为给定问句中每个字符的集合, B 为给定问句中所对应的候选路径中每个字符的集合,则该特征得分的计算公式如下:

$$S_5 = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

(6) S_6 属于路径基础特征得分,表示用户问句与候选路径的词级相似度。对于给定的问句,先使用 jieba 分词库进行分词,从而将句子分成若干词,再使用 Jaccard 计算分词后的问句与路径的词级相似度。

收集上述特征,并基于 XGBoost(极限梯度提升树)算法对以上 6 个特征进行融合计算^[7]。计算结果为每一条候选实体路径对应的特征融合得分,输出范围为 $[0, 1]$ 中的某个值,该值越接近 1,则说明当前候选路径的各项特征越符合问句,即当前候选路径的答案节点作为标准答案的置信度越高;越接近 0 则表示当前路径所对应的答案不能回答用户的提问。

最终,对排名最高的候选路径提取其实体节点和关系,查询知识库,并返回答案节点,回答用户针对火灾应急领域的提问。

3.4 迁移学习策略

由于本文在实体识别、实体链接和路径排序中都使用了涉及亿级别参数量的预训练语言模型,而在火灾应急知识图谱的问答场景中,训练语料中的三元组数量级和问答对仅为百级别,训练时极有可能导致严重的过拟合或预测错误。因此,本文采用迁移学习策略,利用外部语料训练出一个成熟的通用领域知识问答模型,再迁移至火灾应急领域,采用冻结部分网络层的

策略进行模型微调,使得问答模型既具有回复通用领域问题的能力,又能够在火灾应急领域的问题上体现其专业性。外部语料使用的数据集源自 biendata 竞赛平台开展的 CCKS2021 生活服务知识图谱问答评测任务,其中问答语料使用的是由北京大学和北京三快在线科技有限公司人工构建与标注的 8 500 条问答数据,本文将其以 8 : 1 : 1 的比例划分为训练集、验证集和测试集。知识图谱使用的是由开放领域知识库 PKUBASE 及生活服务领域数据集融合得到的知识库数据,本文将其导入 Neo4j 数据集中,构建知识图谱,其中包含 65 535 841 个三元组、16 867 071 个实体项和 408 755 个属性项。

4 实验(Experiment)

4.1 实验数据介绍

本文依据维基百科、百度百科、国家消防救援局、互联网上的其他资料以及相关科普书籍梳理了 544 个三元组、192 个实体、93 个关系,其中包括火灾、火灾事件、可燃物、灭火器、灭火方法、Person(响应人或责任人)6 类概念;InstanceOf、SubclassOf、使用、依附于、参与燃烧、引发火灾、所属类别、调用、选用灭火器 9 种关系类型;中文名、含义、外文名、概念、分类、原理、扑救要点、注意事项、灾害特性、特别警示、特性、级别、结构、适用范围、使用方法、条件、蔓延方式、后果、扑火原则 19 种属性类型。将上述火灾应急领域相关语料整理成 csv 文件储存 Neo4j 数据库中,构建知识图谱。

此外,为方便问答方法的结果验证,本文搜集并整理了 100 个火灾应急领域的问题,人工标注了涉及的实体和三元组路径。示例如下:

问题:A 类火灾的特性?

实体:A 类火灾

三元组:A 类火灾-特性-答案

答案:固体物质火灾,一般在燃烧时能产生灼热的余烬。

4.2 实验结果与分析

本文实验在进行模型训练时,采用 Adam 作为参数优化器,设置其学习率为 0.000 5,Dropout 设置为 0.5,训练最大迭代次数设置为 100 次,批大小参数 Batch Size 设置为 128;在计算模型误差时,选用二元交叉熵损失函数。

实验一:将本文所提方法与传统 Pipeline 方法、传统 Pipeline 进行迁移学习的方法,以及传统 Pipeline 和汪洲等提出的特征融合相结合的方法进行对比实验^[8]。

在火灾应急领域知识图谱问答语料中随机抽取 100 个句子作为测试集语料,评估指标采用准确率 P ,设 Q 为问句集合, $|Q|$ 为问句的个数, TA 为预测答案正确的问句个数,计算公式如下:

$$P = \frac{TA}{|Q|} \quad (3)$$

实验一的结果见表 2。从表 2 中可见,传统 Pipeline 加入迁移学习后方法的准确率为 83.0%,与单独使用传统 Pipeline 方法相比,准确率有了显著的提升,说明本文提出的迁移学习策略可以有效地解决火灾应急领域问答语料较少导致的泛化性差的问题。传统 Pipeline 加入的特征融合方法的准确率为 79.0%,比传统 Pipeline 加入迁移学习后方法的准确率低

4.0%,说明本文的模型在经历迁移学习后更加优秀。本文所提方法与传统 Pipeline 加入迁移学习后方法相比,准确率提升了 6.0%,证明本文所提方法能够提升火灾应急领域知识问答的准确性和可扩展性。

表2 火灾应急领域问答语料的预测结果

Tab. 2 Forecast results of Q&A corpus for fire emergency domain

方法	准确率/%
传统 Pipeline	56.0
传统 Pipeline+迁移学习	83.0
传统 Pipeline+特征融合	79.0
本文方法	89.0

实验二:尝试在 CCKS2021 生活服务知识图谱问答数据集上,将本文所提方法与传统的 Pipeline 进行第一组对比实验;再将本文所提方法与 CCKS2020 新冠百科知识图谱问答评测任务的前两名方案进行第二组对比实验,其中第一名为基于特征融合的中文知识库问答方法^[8],第二名为基于预训练语言模型的检索-匹配式问答方法^[9]。评估指标采用平均 $F1$ 值,设 Q 为问句集合, i 为问句的编号, M_i 为在第 i 个问句中模型预测的答案集合, A_i 为在第 i 个问句中实际为正确的答案集合。计算公式如下:

$$F1_{Avg} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{2|M_i \cap A_i|}{|A_i| + |M_i|} \quad (4)$$

实验二的预测结果见表 3,可以看到本文所提出的问答模型的性能仍显著超越传统 Pipeline 方法,但是与竞赛的先进方案相比还是有一定的差距,分析原因具体如下:第一,CCKS 竞赛中的问句大多都是多跳复杂问句,本文模型在多跳复杂问句的表现上采用了模板扩展的思路,有时未能覆盖 PKUBASE 知识库的全部路径,导致路径遗漏从而输出错误答案。第二,没有考虑包含条件约束的问句的处理方法,例如“北京外环内最近的一家酒店的名称叫什么?”这类问句,蕴含着限制条件和数值表达式,这种限制条件并未纳入本文的考察范畴。

表3 CCKS2021 生活服务问答数据集的预测结果

Tab. 3 Forecast results of CCKS2021 life service Q&A dataset

方法	平均 F1 值/%
传统 Pipeline 方法	62.7
基于特征融合的中文知识库问答方法	86.1
基于预训练语言模型的检索-匹配式问答方法	85.5
本文方法	82.1

5 结论(Conclusion)

本文面向火灾应急知识图谱设计了一种基于传统信息检索式的知识问答 Pipeline 改进方法,首先采用 Task specific 思想将实体识别、实体链接及实体路径排序三个阶段的模型输出结果作为重要特征进行保留,并利用 XGBoost 算法进行特征融合,以避免上、下游任务的误差传递问题;其次引入 Beam Search 算法提升传统路径排序阶段的推理速度,旨在提高问答的搜索效率;最后提出迁移学习策略解决火灾领域小样本模型学习问题。

(下转第 16 页)