基于知识图谱的服装问答系统

赖佳扬,张晓滨,马瑛超
(西安工程大学 计算机科学学院,陕西 西安 710048)

摘 要:随着电商产业的不断发展,消费者希望在网上购物时能够和商家进行更好的沟通,而人工客服需要浪费大量的人力物力。为合理、有效地利用服装资源,文章通过构建服装知识图谱,并基于知识图谱实现服装知识自动问答。该问答系统利用目标实体的多跳关系与问句进行匹配从而完成答案生成,知识问答模型采用 BERT 作为编码层,使用 LSTM 网络对知识库的多跳关系进行学习,利用自注意力机制对知识库特征和问句特征进行计算,最终通过二分类的输出将问句和知识库的匹配结果作为评分,并根据评分给出答案。文章对问答系统的准确性和运行效率在自建的服装数据上进了实验。实验表明,该问答方法相对于传统的答案匹配方法效果更好,同时在运行效率的实验上验证了方法在实际中的可行性。基于知识图谱进行问答系统的搭建可以有效地解答消费者在服装知识和服装搭配推荐上的问题,在提高用户体验的同时节约了人力资源。

关键词:服装;知识图谱;知识问答;知识抽取;Bert 预训练模型
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2023)02-0099-06
doi:10.3969/j.issn.1673-629X.2023.02.015

Clothing Question Answering System Based on Knowledge Graph

LAI Jia-yang,ZHANG Xiao-bin,MA Ying-chao
(School of Computer Science,Xi'an Polytechnic University,Xi'an 710048,China)

Abstract:With the continuous development of the e-commerce industry,consumers hope to have better communication with merchants when shopping online,and manual customer service requires a lot of waste of manpower and material resources. In order to make reasonable and effective use of clothing resources,we construct clothing knowledge graph and realize automatic question answering of clothing knowledge based on the knowledge graph. The question answering system uses the multi-hop relationship of the target entity to match the question sentence to complete the answer generation. The knowledge question answering model uses BERT as the encoding layer,uses the LSTM network to learn the multi-hop relationship of the knowledge base,and uses the self-attention mechanism to calculate the feature of knowledge base and the feature of the question base. Finally,the matching result of the question base and the knowledge base is scored through the binary output,and the answer is given according to the score. The article conducts experiments on the accuracy and operation efficiency of the question answering system on the self-built clothing data. Experiments show that the proposed question answering method has better effect than the traditional answer matching method,and its feasibility in practice is verified in the experiment of operation efficiency. The construction of question answering system based on knowledge graph can effectively answer consumers' questions about clothing knowledge and clothing matching recommendation,which can improve user experience and save human resources.

Key words:clothing;knowledge graph;question answer;knowledge extraction;Bert pre-training model

0 引 言

随着社会的不断发展,消费者的购物方式发生了巨大的改变。服装作为生活中的必需品,已经成为网上购物的第一大商品。由于服装需求的差异性,消费者购物选择时需要频繁地与商家进行沟通。这不但对商家的知识提出了挑战,更重要的是人工服务需要大量的人力资源[1]。知识图谱作为一种结构化的知识库,通过实体关系相互连接,已经成为人工智能应用于知识问答中的重要基础。目前一些大型知识图谱(如 CN-DBPedia)[2-3]已得到了较好的应用。服装作为传统行业,其领域内的智能化技术仍在发展。知识问答作为知识图谱下游应用的重要组成部分,问答方法旨

收稿日期:2022-04-07 **修回日期**:2022-08-10
基金项目:西安工程大学研究生创新基金(chx2021028);国家级大学生创新创业项目(202110709013)
作者简介:赖佳扬(2000-),男,研究方向为知识图谱、自然语言处理;通讯作者:张晓滨(1970-),男,硕士,副教授,研究方向为数据挖掘、个性化服务技术与应用。

在对用户的输入信息进行分析和挖掘,而后利用知识图谱进行答案的搜索,最后反馈给用户。基于知识问答构建智能客服系统能够很好地提高服装智能化水平并减少人力资源需求。

1 研究现状

基于知识图谱的问答方法大致可分为三类:基于语义解析的方法、基于信息检索的方法以及基于知识嵌入的方法。

基于语义解析的方法是利用直接映射或者神经网络方法将自然语言映射成为结构化的表达,而后通过对知识图谱的子图进行匹配完成答案的检索。早期 Steedman 等人[4]提出组合范畴语法,利用词汇表完成问句到结构化表达的转换。孟明明等人[5]提出了一种语义查询拓展方法解决从知识库中无法搜索到理想答案的问题,实现了对知识图谱内容的多语义拓展。上述直接映射的方法依赖于人工制定规则,因此该类方法的适应性很差。Dong 等人[6]基于神经网络进行语义解析,他们通过将映射问题转换为翻译问题,通过一个基于注意力的编码器-解码器结构完成问句的映射。Zhu 等人[7]提出了一种 Tree2seq 模型,该模型将句子映射到知识库的特征空间中以加强其映射的准确性。基于语义解析的方法在问答中是较为规范的流程,这类方法能够有效地减少问答的检索时间,但此类方法的误差对任务的后续影响较大,并影响到最终的结果。

基于知识嵌入的方法则起源于知识图谱推理的发展,包括 TransE[8]、TransR 以及 ITMEA[9]等知识图谱嵌入模型的提出都为知识图谱问答提出了新的思路。Huang 等人[10]提出一种基于嵌入的问答方法,该方法将问题作为输入,并对其进行知识图谱嵌入的映射,在知识图谱的表示中找到其相似度最高的表示,而后将实体作为答案进行输出。Niu 等人[11]将路径和多关系问题间的语义关系引入问答任务,并基于此提出了 PKEEQA 方法,通过实验对比,该方法提高了多关系问答的性能。然而基于嵌入的方法由于需要将图谱和问句进行映射,从而降低了方法的可解释性,这在问答环境下是十分致命的。

基于信息检索的方法则是利用问句中的实体对知识库进行搜索,利用其相关子图进行搜索,进而构成答案集合。Yao 等人[12]利用句法依存树对问句进行解析,分析其问题焦点,而后对知识图谱进行检索,最终通过对比问题焦点、问题核心动词等方法完成答案的确定。Dong 等人[13]利用 Freebase 进行问答系统的构建,该模型利用多列卷积神经网络从答案的路径、背景等对问题进行理解,而后对检索出的答案也使用卷积神经网络进行编码最终比较二者的相似性。Xu 等人[14]在图谱的基础上为实体引入描述信息,进而对描述信息进行编码最终完成答案和问题的匹配。虽然以上方法都取得了较好的结果,但并未考虑到问题涉及实体的多跳关系相互之间的语义联系对答案的影响。

基于此,该文在信息检索的基础上提出一种基于语义匹配的多跳检索问答生成方法。该方法通过在知识库中检索目标实体的多跳信息作为答案集合,而后将生成问题转换为匹配问题,并在模型中设计了以 LSTM 为结构的跳转路径学习层,将学习到的特征与问句进行匹配,然后利用深度学习网络进行匹配评分的计算,最终应用该方法完成了基于知识图谱的服装问答系统。

2 服装知识图谱构建

2.1 模式层构建

模式层又称为本体,其作为知识图谱的骨架,具体规定着知识图谱的实体定义和关系定语。该文采用自顶向下的方法构建服装领域知识图谱,数据以“中国服装网”、“服装网”等专业网站所提供的国家标准 GB/T31007(纺织服装编码)、GB/T 15557-2008(服装术语)为源构建核心概念。实体概念与关系定义如表 1 所示,实体包含概念、设计细节、制造物、生产相关以及教育相关。关系包括上下位概念、组成构成、同义关系以及部分与整体的关系。

如表 1 所示,将服装实体概念进行了五类划分,并为每个分类进行了细分,在实体识别时,模型将按照表中所示的概念进行分类。

表 1 实体概念与关系定义

实体概念定义			实体关系定义		
概念类别	概念标识	二级分类	关系名	语义关系描述	关系实例
概念	Concept	男装、女装、童装、配饰、鞋品	上下位概念	概念术语之间的包含关系	<服装,attribute_of,男装>
设计细节	Design	色彩、风格、款式、细节、制版	组成构成	实体间的组成与构成关系	<羊绒衫,made_of,羊毛>
制造物	Manufactured	面料、辅料、染料	同义关系	同义关系	<丝瓜领,same_of,青果领>
生产相关	Production	生产商、生产工艺	整体与部分关系	部分与整体的关系	<拉链,part_of,夹克>
教育相关	Education	专业名称、教育院校、教材资料			

实体属性在本体中所定义的实体通用属性有中文名称、英文名称、描述信息等,并且参照结构化数据和半结构化数据中的信息为不同的实体类别和小类增加了新的属性信息。如服装面料增加了主要成分、适用衣物、特点、洗涤方式、编织方式、印染方式。

2.2　数据层构建

由于服装领域内网页知识的不规范性,该文使用包装器方式对获取的知识进行进一步的解析,最终获得包含实体和实体关系的三元组。与此同时,为进一步利用服装文本中的知识进行图谱构建,该文对文本信息利用深度学习模型进行知识抽取,这里主要包含对实体的实体识别和关系抽取。

2.2.1　实体识别

实体识别的目的是为了将文本中出现的实体识别出来并确定其类型,该文利用 Bert+BiLSTM+CRF 模型进行服装实体识别。具体的,文本通过 Bert 预训练层[15]将文本进行字符级的编码,而后将编码传入 BiLSTM 网络层进行文本上下文特征的学习,最终传入序列标注层 CRF 得到文本对应字符的识别结果。训练数据使用人工标注的服装实体抽取数据集进行,模型的准确率为 0.94。

2.2.2　关系抽取

关系抽取利用基于输入控制长短期记忆网络模型[16]将实体识别模型识别出的实体进行配对后对其关系进行识别。关系抽取的模型首先通过编码层对文本进行编码,并依据句法依存树生成控制向量,而后利用输入控制长短期记忆网络进行上下文特征的学习,输入控制长短期记忆网络在传统 LSTM 网络上添加输入控制单元,单元结构如图 1 所示。

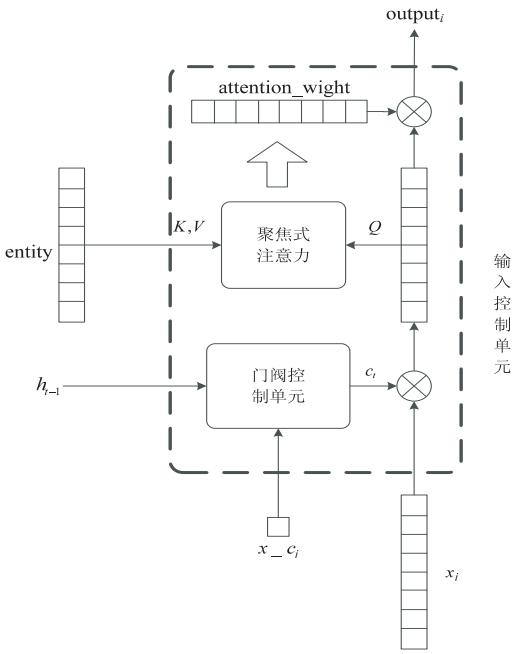

图 1　输入控制单元结构

图中,门阀控制单元计算如下所示:

$$c_t = \sigma(W_c \bullet [h_{t-1}, x_t] + b_c)$$

$$x_t^{'} = x_t \bullet c_t^{x\_c_t}$$

其中,c_t 为输入控制门的控制信息,x_t 为更新后的输入信息,x_c_t 为控制向量,W_c 和 b_c 为记忆门的权重及偏置值。

聚焦式注意力计算进行特征级别的注意力机制的计算,该文选择聚焦式注意力机制将经过门阀控制单元计算出的输入与实体信息进行注意力机制的计算,公式如下:

$$a_i = \text{softmax}(s(x_i, q))$$

其中,a_i 为第 i 个特征的注意力权重,s 为注意力打分函数,x_i 为 t 时刻第 i 个输入特征,q 为 key 的实体特征。

在特征级别的注意力计算中,输入为 t 时刻输入特征的值,q 为实体信息的表示,文本在选择注意力打分函数时使用双线性模型,公式如下:

$$s(x_i, q) = x_i \boldsymbol{W}^{\mathrm{T}} \boldsymbol{q}$$

其中,$\boldsymbol{W}$ 为可训练参数,维度与 $\boldsymbol{q}$ 相同。

最终输入控制单元输出的结果 $\tilde{x}_t$ 为:

$$\tilde{x}_t = \boldsymbol{a} \cdot x_t^{'}$$

其中,$x_t^{'}$ 为注意力模块的输入,$\boldsymbol{a}$ 则代表注意力的权重矩阵。

而后从编码中根据实体位置提取实体特征,进而传入注意力层,将文本特征与实体特征进行注意力计算,最后通过分类网络得到关系抽取的结果。

基于输入控制长短期记忆网络能够对复杂的服装文本进行有效的关系抽取。

2.3　知识构建及存储

以“中国服装网”、“服装网”及百度百科等开源数据进行知识抽取,从中抽取了 5 201 个实体、3 140 条实体关系。图谱囊括了服装材料、服装概念、服装教育相关、设计细节等实体信息。图谱实体分布见表 2。

表 2　图谱实体分布

概念类别	实体数目	示例
概念	1 242	男装、童装、西服
设计细节	1 114	亚麻色、复古风、欧美宫廷风格
制造物	2 128	亚麻、塑料、羊毛
生产相关	219	凯慕琪、漂染、浙江多仕顿
教育相关	498	西安工程大学、东华大学、《服装制版工艺》

通常知识图谱保存的方式有三元组和图数据库,该文的服装知识图谱使用 Neo4j 图数据库进行保存。图 2 展示了部分知识图谱。

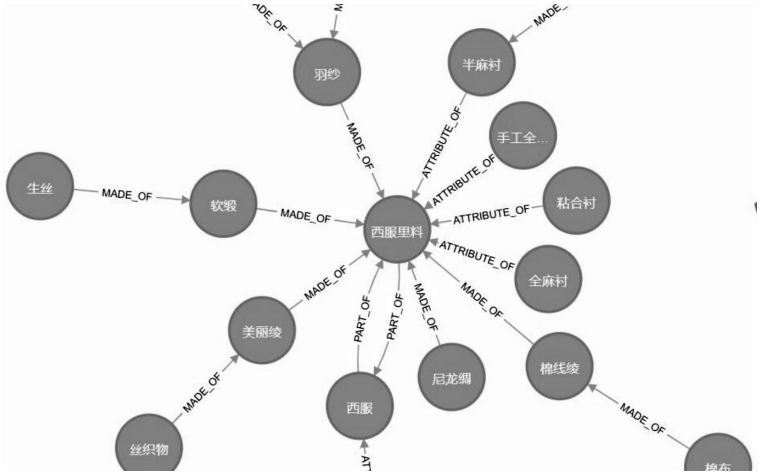

图 2 部分知识图谱

3 基于服装知识图谱的问答系统

3.1 问答系统处理流程

问答系统处理流程包括:输入问题;实体识别;知识库查询;答案匹配;答案回复。

3.2 问答系统的设计实现

3.2.1 服装实体识别

在得到用户的输入后,该文通过命名实体识别模型对问句进行实体识别,采用知识抽取中命名识别模型的 Bert+BiLSTM+CRF 模型,模型结构如图 3 所示。

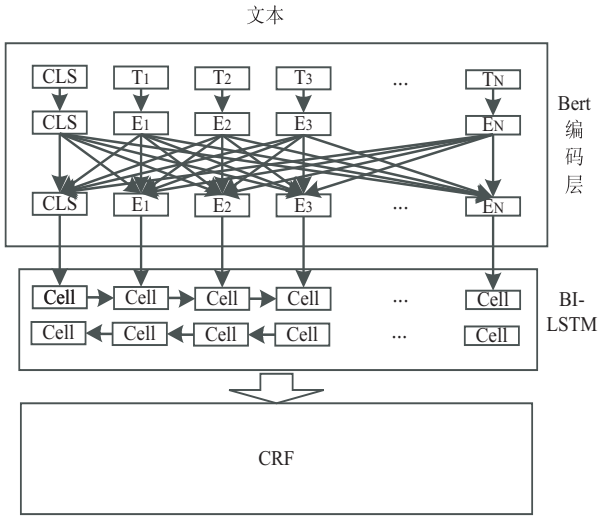

图 3 命名实体识别模型

模型训练采用标注的服装实体识别数据集,训练使用 SGD 优化算法进行参数的调优。

3.2.2 基于 Neo4j 的数据库查询

该文使用 Cypher 语句对 Neo4j 图数据库进行查询,查询语句为:Math(a)-[关系]-(b) where b.name="实体"return a.name。在得到文本中的实体指称后,对实体和查询语句进行拼接后再查询,并且对查询到的实体进行再次搜索,以实现对知识的多跳搜索。实体搜索的范围考虑到问答的一般形式,设置两跳为最大跳转数。最终通过搜索可以得到指称实体围绕两跳范围的所有对象。

3.2.3 答案匹配

在得到实体的多跳数据后,对数据和问句文本进行语义匹配,该文通过建立一个基于 Bert 的语义匹配模型对数据进行匹配,模型结构如图 4 所示。

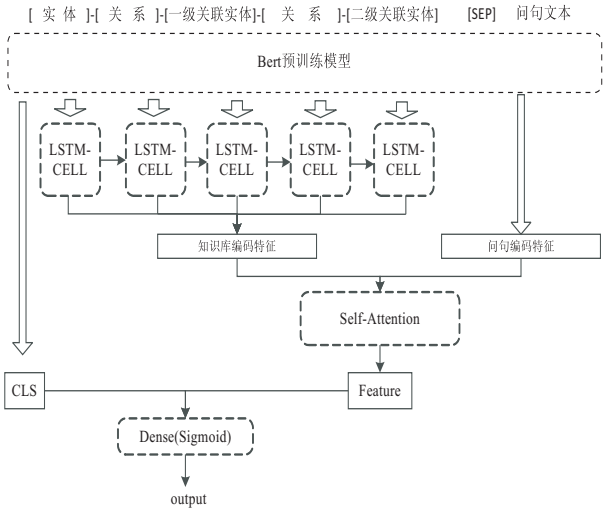

图 4 答案匹配模型结构

模型通过 Bert 编码层对路径和文本进行编码,编码时实体和关系表示均占用十个字符,其余以空格补全。模型利用 Bert 对于拼接文本中用[SEP]拼接后进行编码。利用 LSTM 对搜索到的实体路径进行再次编码,而后与问句编码进行拼接后传入自注意力层,最后使用 Bert 编码层中[CLS]位置的编码与自注意力层拼接后传入以 Sigmoid 为激活函数的全连接层为模型的输出。

模型的优化目标如下:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

其中,N 为样本总数,y_i 为真实样本标签,$\hat{y}_i$ 为模型的预测标签。

模型的输出结果为对于当前搜索和问句的匹配度,最终对匹配度进行排序,选择评分大于阈值的搜索结果作为答案传递给用户。

4　实验结果与分析

为验证该知识问答方法的有效性,并测试基于该知识图谱的服装问答系统的性能,该文对系统答复的准确性和系统的响应时间进行了评测。

4.1　评分阈值选取实验

模型在对文本进行匹配后,系统将评分大于阈值的答案推送给用户,而阈值的选取将直接关系到系统的性能。为选择效果最优的参数,以[0.5,0.9]为值域、以 0.01 为间隔进行了阈值选取实验,实验结果如图 5 所示。

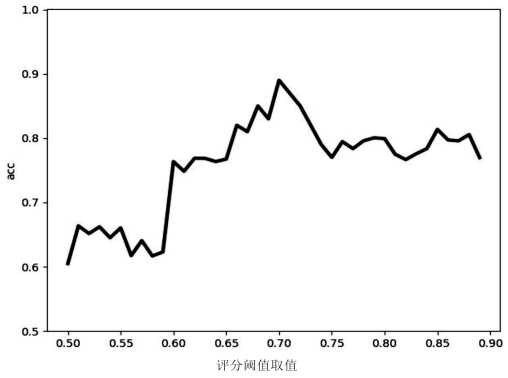

图 5　评分阈值选取实验结果

由图 5 可见,在阈值设置过低时,问答系统会将很多错误的答句反馈给用户,进而导致准确率下降,而当阈值设置过高时,则会过滤掉一些正确答案,因此,系统选择 0.7 作为评分的阈值。

4.2　实验数据及实验环境

该文以自建的服装电商领域服装知识和服装搭配推荐问答语料集进行模型训练和验证。数据来源为“中国服装网”中的问答栏目内容和与学校合作电商的问答数据记录,将 6 325 条数据作为训练数据,2 000 条作为验证数据。

模型训练硬件环境为 RTX3060 及 Intel(R) Core(TM) i7 - 10750H 。软件环境为 python3.6,tensorflow1.14.0,keras2.2.4,Bert 版本为 Bert-chinese-base。系统部署于阿里云服务器,后端框架为 SSM。

4.3　答案准确性实验

4.3.1　实验评测指标

实验评价指标选用精确率 P 。给定输入问答集 Q ,对于 Q 中每条问句 q 由 N 个答案组成。则问答精确率的评价指标定义如下:

$$P = \frac{\sum_{q \in Q} |E_q \cap E_q^{'}|}{\sum_{q \in Q} |E_q^{'}|}$$

其中,问答对应的答案实体编号为 $E_q = \{e_1, e_2, \cdots, e_n\}$;模型输出的答案结果为 $E_q^{'} = \{e_1^{'}, e_2^{'}, \cdots, e_n^{'}\}$ 。

4.3.2　实验结果

为验证该问答系统的有效性,以自建的服装电商领域服装知识和服装搭配推荐问答语料集进行实验。将传统的模板匹配方法(Template Match)、图谱嵌入方法(TransE、TransR)和语义解析方法(NER+RE、Tree2seq)和文中方法进行对比,实验结果如表 3 所示。

表 3　问答方法准确性实验

方法名	准确率
Template Match	0.74
TransE	0.86
TransR	0.81
NER+RE	0.77
Tree2seq	0.79
文中方法	0.89

由表 3 可见,基于模板匹配的方法相对于文中方法在准确率上有所不足,原因是模板匹配方法依赖于模板的制定,而模板的制定需要付出巨大的代价以适应各个问答语句。而基于知识图谱推断的方法相对传统方法取得了很好的效果,但基于知识图谱推断的方法缺乏可解释性且在稀疏图谱中效果不佳。而文中方法通过准确率较高的实体识别方法将两跳以内的实体进行搜索,而后通过语义匹配方法进行答案生成,这相对于使用实体识别和关系抽取先判别意图和实体而后搜索的方法具有更高的准确性。

4.4　系统响应时间实验

为验证系统的适用性,该文使用无答案、一跳查找以及两跳查找的问句进行测试,测试结果如表 4 所示。

表 4　系统响应时间实验结果

测试数据	平均响应时间/s
无答案	3.1
一跳搜索	2.4
两跳搜索	2.6

文中方法由于存在命名实体识别和匹配两个阶段,且使用了 Bert 预训练模型,因此系统的响应速度相对较慢,但响应时间总体符合系统的应用环境。

5　结束语

为了满足服装电商领域智能客服问答的需求,设计实现了以服装知识图谱为基础的服装知识问答系统。该系统基于信息检索的方式进行答句生成,提出的答句生成方法通过在知识库中检索目标实体的多跳信息作为答案集合,而后将生成问题转换为匹配问题,

并设计了以 LSTM 为结构的学习跳转路径。最后的实验验证了该问答系统的有效性,可满足服装领域智能客服环境中服装知识和服装搭配推荐问答的需求。

参考文献:

[1] 田 苗,李 俊. 智能服装的设计模式与发展趋势[J]. 纺织学报,2014,35(2):109-115.

[2] SHI W L, WANG Z H, CHEN H, et al. Exploring the core knowledge of business intelligence[C]//23rd Pacific Asia conference on information systems 2019. Xi'an:[s. n.],2019:41.

[3] LEHMANN J,ISELE R,JAKOB M,et al. DBpedia - a large-scale,multilingual knowledge base extracted from Wikipedia[J]. Semantic Web,2015,6(2):167-195.

[4] STEEDMAN M. Surface structure and interpretation[J]. Linguistic Inquiry Monographs,1996,43(3):271-276.

[5] 孟明明,张 坤,论 兵,等. 一种面向知识图谱问答的语义查询扩展方法[J]. 计算机工程,2019,45(9):276-283.

[6] DONG L,LAPATA M. Language to logical form with neural attention[J]. Office for Official Publications of the European Communities,2016,1:33-43.

[7] ZHU S,CHENG X,SU S. Knowledge-based question answering by tree-to-sequence learning[J]. Neurocomputing,2020,372:64-72.

[8] ANTOINE B,NICOLAS U,JASON W. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems 26:27th annual conference on neural information processing systems 2013. Lake Tahoe:[s. n.],2013:2787-2795.

[9] WANG Z,ZHANG J W,FENG J L,et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the twenty-eighth conference on artificial intelligence. [s. l.]:AAAI,2014:1112-1119.

[10] HUANG X,ZHANG J Y,LI D C,et al. Knowledge graph embedding based question answering[C]//Proceedings of the twelfth international conference on web search and data mining. Melbourne:ACM,2019:105-113.

[11] NIU G,LI Y,TANG C,et al. Path-enhanced multi-relational question answering with knowledge graph embeddings[J]. arXiv:2110.15622v1,2021.

[12] YAO X,DURME B V. Information extraction over structured data:question answering with freebase[C]//Proceedings of the 52nd annual meeting of the association for computational linguistics. Baltimore:ACL,2014:956-966.

[13] LI D,WEI F,MING Z,et al. Question answering over freebase with multi-column convolutional neural networks[C]//International joint conference on natural language processing(ACL-IJCNLP). Beijing:ACL,2015:260-269.

[14] XU Y,ZHU C,XU R,et al. Fusing context into knowledge graph for commonsense question answering[C]//Annual meeting of the association for computational linguistics(ACL). [s. l.]:ACL,2021:1201-1207.

[15] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics:human language technologies. Minneapolis:ACL,2019:4171-4186.

[16] 马瑛超,张晓滨. 基于控制输入长短期记忆网络的关系抽取方法[J]. 计算机系统应用,2022,31(3):282-287.

(上接第 98 页)

[11] GAILLY J L,ADLER M. ZLIB documentation and sources [EB/OL]. 2020. ftp://ftp.uu.net/pub/archiving/zip/doc/. 2020,9.

[12] DRUET A,GOSSET N,ROBINSON J A. Binary-tree recursive motion estimation for video coding[C]//1997 sixth international conference on image processing and its applications. Dublin:IET,1997:51-55.

[13] SULLIVAN G J,OHM J,HAN W,et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology,2012,22(12):1649-1668.

[14] MUKHERJEE D. A technical overview of VP9 - the latest open-source video codec[J]. SMPTE Motion Imaging Journal,2015,124(1):44-54.

[15] JOSHI K,YADAV R,ALLWADHI S. PSNR and MSE based investigation of LSB [C]//2016 international conference on computational techniques in information and communication technologies (ICCTICT). New Delhi:IEEE,2016:280-285.

[16] ARNOLD R,BELL T. The canterbury corpus[EB/OL]. 2020. https://corpus.canterbury.ac.nz/descriptions/. 2020,9.

[17] DEOROWICZ S. Silesia compression corpus[EB/OL]. 2020. http://sun.aei.polsl.pl/~sdeor/index.php? page=silesia. 2020,9.

[18] ARNOLD R,BELL T. A corpus for the evaluation of lossless compression algorithms[C]//Proceedings DCC '97. data compression conference. Snowbird:IEEE,1997:201-210.