基于BERT-CNN的电影原声智能问答系统

黄东晋,秦 汉,郭 昊
(上海大学 上海电影学院,上海 200072)

摘 要:智能问答是自然语言处理领域一个非常重要的研究热点,传统的智能问答不能准确地理解用户的意图,从而无法返回准确的答案。因此,提出了基于BERT-CNN算法的智能问答系统,并应用于电影原声领域,可以快速准确地反馈相关信息。首先,构建电影原声的知识图谱,建立节点实体以及实体之间的关系,利用Neo4j图数据库对数据进行存储。然后,通过基于规则和词典的方法进行实体识别,利用BERT-CNN分类算法对用户意图进行分类。最后,根据用户意图和实体,将问句转化成知识图谱的查询语句,在数据库中查询后返回结果。实验结果表明,构建的面向电影原声智能问答系统是可行的,采用BERT-CNN分类算法,分类准确率高达91.24%,能够实时得到问题答案的准确反馈,准确率达到95%以上。

关键词:智能问答;知识图谱;电影原声;BERT-CNN分类;图数据库
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2020)11-0158-05
doi:10.3969/j.issn.1673-629X.2020.11.029

Movie Soundtrack Intelligent Question and Answer System Based on BERT-CNN

HUANG Dong-jin, QIN Han, GUO Hao
(Shanghai Film Academy, Shanghai University, Shanghai 200072, China)

Abstract: Intelligent question answering is a quite important research hotspot in the field of natural language processing. Traditional intelligent question answering cannot accurately understand the user's intention, so it cannot return accurate answers. Therefore, an intelligent question-and-answer system based on BERT-CNN algorithm is proposed and applied to the field of movie soundtrack, which can feedback relevant information quickly and accurately. First, the knowledge map of movie soundtrack is constructed, node entities and relationship between entities are established, and Neo4j graph database is used to store the data. Then, entity recognition is carried out based on rules and dictionaries, and user intention is classified by BERT-CNN classification algorithm. Finally, according to user intention and entity, the question is transformed into the query statement of knowledge graph, and the result is returned after the query in the database. The experiment shows that the constructed intelligent question-and-answer system oriented to the film soundtrack is feasible. The BERT-CNN classification algorithm is adopted, with the classification accuracy as high as 91.24%, and the accurate feedback of the questions can be obtained in real time, with the accuracy of more than 95%.

Key words: intelligent question answering system; knowledge graph; movie soundtrack; BERT-CNN classification; graph database

0 引 言

随着移动互联网的飞速发展,在大数据时代,用户可以上网通过谷歌、百度等引擎搜索自己想要浏览的信息。搜索引擎极大地方便了人们的日常生活,但是也存在不少问题,它没法精确地返回用户所需的答案,用户只能浏览相关的网页去找寻自己的答案,这样浪费了用户的时间和精力。而且,搜索引擎是通过关键字来返回相关网页信息,没法理解用户输入的问题中所含的语义信息。智能问答系统(intelligent question answering system)已经有半个多世纪的研究历史,互联网的飞速发展使得基于知识图谱的问答系统成为当下学术界以及互联网公司研究的热点。基于知识图谱的问答系统可以根据用户输入的问句直接返回精确的答案,它不需要用户从海量的相关信息中找到自己想要的答案,节约了用户的时间和精力。

基于知识图谱的问答系统也被简称为知识库问答系统。主要分传统的知识库问答和基于深度学习的知识库问答。传统的知识库问答又包括基于语义解析和

收稿日期:2020-01-08 **修回日期**:2020-05-11
基金项目:国家自然科学基金(61402278);上海市自然科学基金(19ZR1419100);上海大学电影学高峰学科项目(19ZR1419100)
作者简介:黄东晋(1982-),男,博士,副教授,研究方向为计算机图形学、虚拟现实、影视技术等;秦 汉(1995-),男,硕士研究生,研究方向为自然语言处理。

基于信息抽取的知识库问答。语义解析知识库问答是从语义上分析自然语言,将其转化为知识图谱可以“理解”的形式,然后在图谱中进行查找答案。语义解析的方法通常会用到依存分析[1]或组合文法[2]的方法来分析和处理自然语言,然后转化成知识图谱查询语言查找答案。Berant 等人[3]使用了基于语义解析的知识库问答文章。基于信息抽取的知识库问答则是抽取问题的关键信息,此类方法去除了繁琐的语义解析。Yao 等人[4]通过分析自然语言抽取出问题里面的主题实体,然后通过分类器训练从一系列候选答案中选出正确答案。区别于传统的知识库问答,基于深度学习的知识库问答使用表示学习[5]的技术,将问题和答案都用向量来表示,从而问答任务变成问题和答案之间匹配问答问题[6]。Dong 等人[7]采用三个文本卷积网络(text-CNN)分别从答案的三个方面进行表示学习,这三个方面为答案类型、答案路径、答案上下文。Yin 等人[8]在处理问题生成的短语结构树时引入了递归神经网络,递归神经网络通过分析问题中的隐藏关系和类型意图来提高回答的准确率。Yih 等人[9-10]采用卷积神经网络语义模型和分段查询图生成,开发了基于语义相似度的语义解析框架。Hao 等人[11]提出了一种使用双向 LSTM 并结合问题引入注意力机制提高问句特征的知识图谱问答设计。近年来市场上也出现了很多问答机器人,例如苹果的 siri、微软的小冰等,但这些机器人的知识库大多偏向开放领域的,针对具体专业的问题,可能回答不出来。目前知识图谱在电影原声方面的应用很少,无法满足用户对电影里面电影原声的曲目、发行时间以及出版社等问题的查询。因此,该文基于 BERT-CNN 分类算法构建了一个面向电影原声的智能问答系统,可以准确地回答电影原声相关的问题,并且采用 BERT-CNN 算法进行问题分类提高问题分类的准确率,在答案查询部分增加相似性查询也提高了系统的实用性。

1　系统框架

该文设计的基于知识图谱的问答系统,主要结合知识图谱将用户输入的自然语言问句转化为 Cypher 查询语句。首先需要对用户输入的自然语言问句进行预处理,比如分词、去停用词等,得到实体,然后根据分类算法可以获得问题的分类类别,从而找到该类别的查询模板,接着将实体填入查询模板中就可以从知识图谱中获得用户所需的答案。该系统分为三部分,分别为知识图谱构建、问题预处理和答案生成。系统框架如图 1 所示。

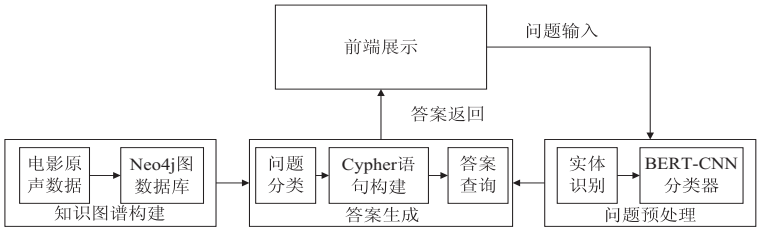

图 1　系统框架

· 知识图谱构建部分:主要是将网上爬取的数据进行结构化处理,然后以<节点,关系,节点>三元组的形式存入数据库中。

· 问题预处理部分:将用户输入的问题通过 jieba 分词,去停用词等,进行词性标注,然后通过基于规则和词典的方法识别出实体,最后将实体与属性送到 BERT-CNN 分类器中训练,得到问题的分类结果。

· 答案生成部分:根据分类结果选择不同的 Cypher 查询模板,然后构造相应的查询语句从知识图谱中获得用户所需的答案。

2　电影原声知识图谱

知识图谱[12]是一个拥有很多节点和关系的语义网络,它可以把与关键词相关的知识体系系统化地展示给用户。例如文中构建的电影原声方面的知识图谱,既可以使用户更加全面准确地了解电影原声的信息,又可以使用户花费更少的时间拥有更佳的关于电影原声信息的搜索质量和搜索体验。

2.1　数据的获取

该文选择爬豆瓣有关电影原声的相关网页,为了构建电影原声知识图谱中的各类节点,在网页中爬取“片名”、“表演者”、“流派”、“介质”、“发行日期”、“出版社”、“相关电影”、“曲目”、“评分”、“简介”标签内的内容,然后将爬取到的内容进行进一步处理,最终以 Json 文件的格式进行保存。整个数据库大约有 1 000 多条电影原声信息,关于最新的电影原声信息,也会通过手动添加数据集的方式使知识图谱趋于完备。

2.2　知识图谱的构建

知识图谱是通过定义实体与实体间的关系来把知识串起来的,“实体”是知识图谱中节点的表示方法,图谱中的边是由“关系”来描述的。以电影原声的知识图谱为例,实体是电影原声,出版社和曲目等,而关系则是用来连接两个实体的,例如电影原声和曲目的关系是“包含”。

该文选择 Neo4j 图数据库[13]来存储数据,构建知识图谱。Neo4j 是 NoSQL 的图形数据库,性能比较高。它可以把结构化的数据存储在网络上而不是表格中,非常适合知识图谱的存储方式。由于任意的节点都能通过关系与其他节点相连接,节点也可以有很多的属性,所以也使得添加数据变得非常方便。

接下来将展示电影原声的知识图谱,如图 2 所示,这里将电影原声作为一个中心节点,此节点中包含电影原声的相关属性,例如流派、介质、相关电影、评分以及歌手等等。从这个中心节点也能看到与其他节点的关系,例如电影原声和出版社的关系是“press”。

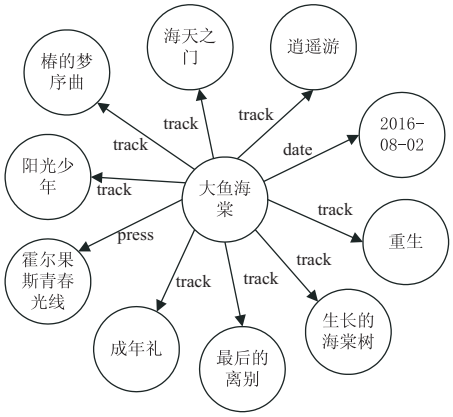

图 2 电影原声知识图谱

3 答案生成算法

3.1 问句分类

在问句分类环节中,首先需要按照问句设计模块确定查询意图。该文将问题意图分为 10 类,然后每一类中再以不同的问法构建数据集。最终通过手动构建的训练集约 20 000 条,测试集和验证集大约也都有 2 000条。把 BERT[14]当作 embedding 层送入 CNN[15]模型中可以取得更好的分类准确率,分类准确率达到 91.24%,所以选择 BERT-CNN 分类算法来做问句的分类,其他分类算法的准确率如表 2 所示。

BERT:BERT 模型是一个多层双向 Transformer 编码器,Transformer 是一种注意力机制,可以学习到文本中单词的上下文关系。Transformer 原型包含 encoder 机制和 decoder 机制,encoder 作为输入接受文本,decoder 主要负责预测结果。BERT 的出现使得预训练产生词向量与下游具体 NLP 任务(对词向量的操作)的关系发生了很大的改变,从 word2vec 到 BERT,主要工作是把下游具体 NLP 任务的工作慢慢转移到预训练词向量上。word2vec 是基于单词级别的,它的缺点是上下文无关,而 BERT 是基于句子级别的,Transformer 做 encoder 可以实现上下文相关,BERT 模型进一步提高了词向量模型的泛化能力,可以充分描述字符级,词级以及句子级的特征。BERT 的模型如图 3 所示,其中 $E_1 - E_N$ 为一个句子的词嵌入,Trm 为 Transformer 的编码器结构[16],Transformer 不需要循环,而是并行处理序列中的所有单词或符号,同时利用自注意力机制将上下文与较远的单词结合起来,充分考虑了上下文信息。$T_1 - T_N$ 为输出,分别对应其上下文。

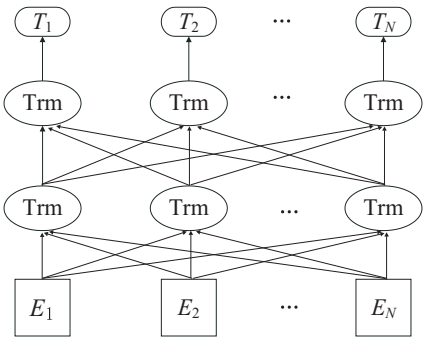

图 3 BERT 模型

卷积神经网络(convolutional neural networks, CNN):最初应用于图像处理,并且在图像处理领域取得了非常好的效果,同时它可以应用在文本分类上面。文本分类主要是准确无误地提取句子或文档的中心思想,将句子或文档的关键字作为特征去训练分类器进行分类,而 CNN 的卷积和池化就是一个抽取特征的过程。CNN 模型如图 4 所示,通常包括卷积层、池化层和全连接层。卷积层的作用是进行特征提取,然后将其特征图作为池化层信息过滤和特征选择的输入。池化层的作用是降低维度,通过降采样可以更进一步降低维度,池化层分为最大池化和平均池化两大类。CNN 中的全连接层就相当于传统前馈神经网络中的隐含层,它位于 CNN 隐含层的最后部分,并且只向其他全连接层传递信号。

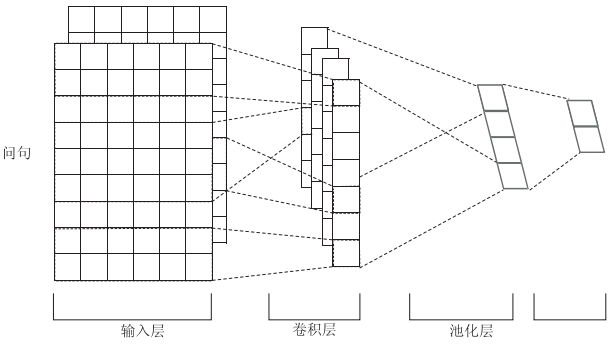

图 4 卷积神经网络示意图

令 $X_i \in R^k$ 表示句子中第 i 个字的 k 维向量,那么长度 n 为句子(不够要补齐到 n),则可以表示为:

$$X_{1:n} = X_1 \oplus X_2 \oplus \cdots \oplus X_n \tag{1}$$

其中,$\oplus$ 是连接符。令 $\omega \in R^{hk}$ 表示卷积核,它的窗口大小为 $h * k$,那么通过卷积操作的特征 c 为:

$$c_i = f(\omega \cdot X_{i:i+h-1} + b) \tag{2}$$

其中,$b \in R$ 是偏执变量,f 是非线性激活函数,通过卷积操作,句子 $X_{1:n}$ 转化成特征图 c:

$$c = [c_1, c_2, \cdots, c_{n-h+1}] \tag{3}$$

接着通过最大池化的方式,将特征图 c 变成 $\hat{c} = \max\{c\}$,一个卷积核提取特征的操作是用每个特征图中值最大的特征来表示整个特征图,CNN 通常会使用多个窗口大小不同的卷积核来提取多个特征,最后在全连接 Softmax 层根据特征进行分类。

3.2　答案查询

3.2.1　全匹配查询

Cypher 语句是 Neo4j 图数据库的查询语言,它与 SQL 语句类似,内容丰富,同时也包含着很多封装好的函数。该文通过命名实体识别和问题分类,就可以把用户输入的问句转换成 Cypher 查询语句,然后在 Neo4j 中执行就可以得到用户所需要的答案。该文针对不同类型的问题制定了不同的 Cypher 查询模板,例如:

(1)查询“电影原声发行时间”的模板,已知实体“电影原声(Music)”,还有两个实体之间的关系,这里是电影原声与发行时间的关系:“date”,就可以得到实体“发行时间(Date)”:MATCH (m:Music)-[r:date]->(n:Date) where m.name='{0}' return n.name。

(2)查询“电影原声评分”的模板,已知实体“电影原声(Music)”,就可以获得属性“评分”的值:MATCH (m:Music) where m.name='{0}' return m.name, m.score。

3.2.2　相似度匹配查询

用户在进行自然语言输入时,难免会遇到打字错误或者增删了某些字,这样采用全匹配时就无法从字典中匹配到相应的实体,从而在 Neo4j 中返回正确的答案。为了增强问答系统的实用性,该文还提供了第二种查询即相似度查询。它是通过计算词语和字典中的词的相似度,如果最终得分大于 0.7,则说明它们具有相似性,这里的得分是余弦相似度评分和编辑距离评分的均值。例如误把“红楼梦”输入成“红楼”时,即当用户输入“红楼的评分是多少”,系统首先采用全匹配发现找不到实体“红楼”,接着就采用相似度匹配发现和“红楼梦”的相似度评分达到 0.736,超过阈值 0.7,最终返回答案“电影原声红楼梦的评分是:9.7”。

4　实验结果与分析

该系统可以对用户的问题进行实时分析并给出准确的答案,系统的硬件环境是:Inter(R) Core(TM)i5-8300H CPU @ 2.30 GHz 的处理器,8 GB 的内存,NVIDIA GeForce GTX 1050 Ti 的显卡。该系统是基于电影原声知识图谱的,首先把用户输入的问句预处理识别出实体,然后通过 BERT-CNN 算法将问题进行分类。实验结果表明,在 10 类问句的数据集上,分类的准确率高达 91.24%。最后进行答案查询,答案查询时首先进行全匹配查询,全匹配失败时采用相似度查询,从而得到问题的答案。实验结果如表 1 所示。

表 1　实验结果

问题	问题分类	答案
你的名字的评分是多少?	评分	电影原声你的名字评分是 9.3
大鱼海棠的发行公司是?	出版社	电影原声大鱼海棠的出版社是霍尔果斯青春光线
你的名字的是什么时候发行的?	发行时间	电影原声你的名字的发行时间是 2016-09-02

在用户问题分类时采用了多种算法进行比较,最终基于该文自己构建的数据集进行实验后发现 BERT-CNN 的分类效果最好,达到 91.24%,其次是 BERT,分类准确率达到 89.98%,效果最差的是 NB(朴素贝叶斯算法),准确率只有 80.89%。所以选用 BERT-CNN 分类算法来进行问题分类,实验结果如表 2 所示。

表 2　分类算法准确率比较

分类算法	准确率/%
BERT	89.98
CNN	87.79
BERT-CNN	91.24
NB	80.89

将该智能问答系统与一个使用协同过滤的电影推荐系统集成到浏览器上运行,可以兼容多种浏览器,开发技术选择了 Flask 微型 Web 开发框架。系统整体流程是首先进行用户登录界面,基于用户协同过滤的电影推荐系统可以针对不同的登录用户推荐 4 部电影,然后智能问答系统又可以回答用户关于电影原声方面的问题。效果如图 5 所示,首先问答规则说明用户可以按照这样的类别来进行问题的输入,然后用户在输入框中输入问题后点击搜索就可以获得问题的答案了。

5　结束语

提出了一种基于 BERT-CNN 的分类算法,并且把该算法应用到智能问答系统的用户意图查询部分。该智能问答系统可以准确回答电影原声领域的相关知识,BERT-CNN 分类算法可以精确地查询用户问题的意图,答案查询部分增添相似度查询也有效提高了系统的实用性。下一步工作是改进问题预处理模块,基于自定义词典的实体识别没有捕捉到词与词之间的语义关系,并且还要耗费时间来构建词典,因此可以采用深度学习的方式来获得更好的识别效果。

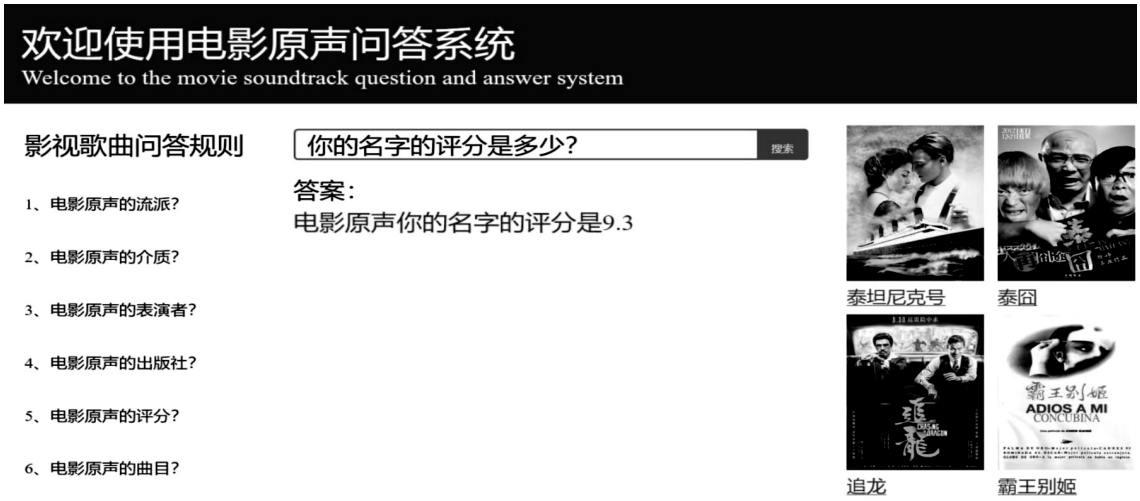

图 5 问答系统

参考文献:

[1] LIANG P,JORDAN M I,KLEIN D,et al. Learning dependency-based compositional semantics[J]. Computational Linguistics,2011,39(2):389-446.

[2] CAI Q,YATES A. Large-scale semantic parsing via schema matching and lexicon extension[C]//Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers). Sofia, Bulgaria: ACL, 2013: 423-433.

[3] BERANT J,CHOU A,FROSTIG R,et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. Seattle: ACL,2013:1533-1544.

[4] YAO X,VAN DURME B. Information extraction over structured data: question answering with freebase[C]//Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers). Baltimore, Maryland: ACL,2014:956-966.

[5] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(2):247-261.

[6] YIN W,YU M,XIANG B,et al. Simple question answering by attentive convolutional neural network[C]//Proceedings of COLING 2016,the 26th international conference on computational linguistics: technical papers. Osaka, Japan: The COLING 2016 Organizing Committee,2016:1746-1756.

[7] DONG L,WEI F,ZHOU M,et al. Question answering over freebase with multi-column convolutional neural networks [C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Beijing: ACL,2015:260-269.

[8] YIN Jun,ZHAO Wayne Xin,LI Xiaoming et al. Type-aware question answering over knowledge base with attention-based tree-structured neural networks[J]. Journal of Computer Science and Technology,2017,32(4):805-813.

[9] YIH W,HE X,MEEK C. Semantic parsing for single-relation question answering[C]//Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers). Baltimore, Maryland: ACL,2014: 643-648.

[10] YIH W,CHANG M W,HE X,et al. Semantic parsing via staged query graph generation: question answering with knowledge base[C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing(volume1: long papers). Beijing: ACL,2015:1321-1331.

[11] HAO Y,ZHANG Y,LIU K,et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). Vancouver, Canada: ACL, 2017:221-231.

[12] 刘 峤,李 杨,段 宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展,2016,53(3):582-600.

[13] STONEBRAKER M. SQL databases v. NoSQL databases [J]. Communications of the ACM,2010,53(4):10-11.

[14] DEVLIN J,CHANG M,LEE K,et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//The 18th annual conference of the north American chapter of the association for computational linguistics: human language technologies. New Orleans, Louisiana: ACL, 2019:4171-4186.

[15] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha, Qatar: ACL,2014:1746-1751.

[16] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Advances in neural information processing systems. Long Beach: NIPS,2017:5998-6008.