

文章编号: 2096-1618(2019)03-0232-06

基于深度学习的领域问答系统的设计与实现

胡 婕, 陶宏才

(西南交通大学信息科学与技术学院, 四川 成都 611756)

摘要:智能问答系统可以快速、准确地为用户提供信息服务,是目前自然语言处理领域一个备受关注的研究方向。旨在结合深度学习算法实现一个基于知识库的领域问答系统。在知识库构建上,首先采用 protégé 编辑本体模型,在完成领域知识获取以及 RDF 知识图谱构建后,以 Fuseki Server 作为 SPARQL 服务器向问答系统检索模块提供查询接口。在问句意图识别方面,通过多种分类算法在领域问题集上的对比实验,选用深度学习中的 CNN 模型用于问句分类。在槽位获取上,灵活地将序列标注模型 Bi-LSTM+CRF 应用于问句中的槽位提取。测试结果表明,实现以微信公众号为问答终端的电影领域问答系统,在问句集上的问句理解准确率达到 98%;基于构建的领域知识图谱,可以比较准确迅速地回答大部分领域问题。

关键词:问答系统;深度学习;知识图谱;短问句分类

中图分类号:TP312

文献标志码:A

doi:10.16836/j.cnki.jcuit.2019.03.004

0 引言

随着大数据时代的到来,针对信息爆炸和信息过载问题的搜索引擎只返回一系列相关网页的信息检索方式已经无法满足需求,而问答系统(question answering system, QAS)则在一定程度上弥补了搜索引擎的不足。问答系统能让用户以自然语言直接提问,同时可以返回给用户精确、简洁的答案,极大地提高了获取信息的效率,也因此成为目前自然领域处理中最热门的研究方向^[1-2]。问答系统按领域可以分为开放领域和限定领域。目前,虽然有一些比较成熟的开放领域的问答系统,如 Siri、微软小冰等等,但是由于开放领域问题范围太广,开放领域问答系统对用户的问题只能做浅层分析来回答,有时甚至会出现回答的内容与问题不相关的情况^[3]。而限定领域的问答系统可以借助很多领域相关资源,从而极大地提高了系统的可用性 & 实用性。

近年来,深度学习技术的不断发展,为智能问答领域指明了新的发展方向。然而目前深度学习算法在智能问答系统中的应用还不够充分。此外,各大商业公司对于自然语言问答系统的研究都还比较封闭^[4]。

为此,在研究限定领域问答系统的同时对深度学习的相关算法进行研究,进而将深度学习算法应用于智能问答系统中,实现了一个基于领域知识库的融合深度学习算法的问答系统。

1 智能问答系统的研究现状

问答系统的起源可以追溯到 20 世纪 60 年代。1961 年, BASEBALL^[5] 就可以回答美国棒球赛的时间地点成绩等问题。1966 年,出现了第一个 QA 系统 Eliza^[6]。Eliza 扮演心理学专家,采用基于模式和关键字匹配的方法,将用户的问题转换为查询语句,进而获取出答案返回给用户。由于其针对特定应用领域,因此可以事先构造领域模型,使用推理技术来回答用户的复杂问题。不过,构建这类系统的模式库和规则库十分困难,而且需要领域专家的综合领域知识^[7]。

自 1999 年 TREC 举办首届自动问答竞赛之后(类似的问答系统评测竞赛还有 CLEF 和 NTCIR),研究者们开始从信息检索的角度来开展问答系统的研究工作,研究领域也逐渐从最初的限定领域拓展到开放领域,研究对象也从固定语料向互联网扩展,参加这些竞赛的各类问答系统的性能也在逐年提高。在此阶段,问答系统主要是通过搜索关键字得到相关文档,再根据问题的核心语义,从文档中找到包含答案的文本片段,最后抽取出问题的答案。这类开放领域的问答系统一般比较容易评测,但是通常只对问句作浅层次的分析^[8],代表系统有 AskJeeves^[9], START^[10]。

针对上述不足,很多互联网公司开始研发面向社区的问答,如雅虎、百度知道、知乎和搜狗问问等。这种问答系统通过维护一个大的问答社区,将与用户问题相似的历史问题及历史答案返回给用户。这种问答

模式的核心是基于关键字匹配的技术,不需要经过问题理解、信息检索和答案抽取等复杂的处理过程^[11]。

2010 年左右,谷歌提出了知识图谱的概念。通过把一些非结构化的纯文本数据变成结构化数据,然后基于结构化的知识图谱来做问答服务,称之为 KB-based QA,也就是基于知识图谱的问答系统^[12]。这类问答系统的核心是如何基于已有的数据库或知识图谱,准确地理解用户的问题,并找到对应的答案。

到目前为止,主流问答机器人大多都是面向开放领域的。比如,用海量聊天语料为基础的微软小冰和苹果的 Siri,以及基于用户搜索日志的百度语言助手等。然而,开放领域问答系统对一些特定领域(如旅游、医学和客服等)的问题无法做到较好的回答。同时,由于中文表达字间无分隔的特性,不像英文表达那样具有天然的分词特性,因此使得基于中文的自然语言处理问答系统的研究更为困难^[13]。

2 系统设计

研究目标为实现一个基于中文自然语言处理的限定领域的问答系统,领域限定为电影。该系统可完成对电影领域自然语言问题的理解,并在充分理解用户意图后,将用户的自然语言查询转换为知识库可理解的结构化查询语句,最后检索查询并返回给用户简洁的答案。

2.1 系统模块设计

针对研究目标,该电影问答系统主要可以分为 4 大模块:

(1) 知识库建立模块:要建立一个基于知识图谱的问答系统,首要任务就是针对选取的领域构建一个规模较大的知识库。知识库是该类问答系统的基础,知识的数量与质量是一个基于知识库的问答系统性能是否优越的决定性因素。

(2) 问句理解模块:是问答系统的核心,只有准确地理解了用户意图和关键信息,后续的知识检索过程才是有效的。

(3) 问句转换模块:将问句理解模块输出的用户意图和关键槽位(即句子中的关键信息)进行转换,生成知识库能够理解的结构化查询语言。

(4) 答案检索模块:利用问句转化模块生成的知识库查询语句,向知识库发起检索,得到问题的答案。

2.2 系统框架设计

针对电影领域问答系统的需求,构建的系统架构

如图 1 所示。

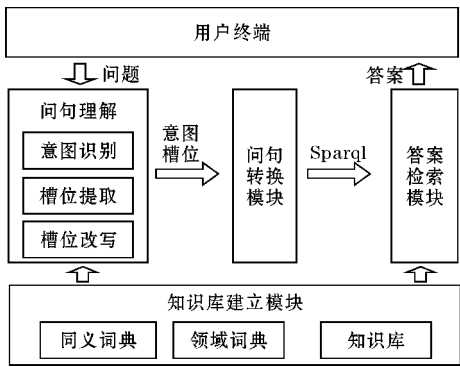


图 1 系统架构图

知识库是基于知识图谱的问答系统(knowledge base question answering, KB-based QA)的基础。知识库建立模块负责从豆瓣电影网站获取电影领域知识,并将知识抽取成三元组的形式,以此形成该领域的知识图谱。同时,利用领域知识构建同义词典和领域词典(电影名和人名词典等)为其他模块提供服务。

在拥有了领域知识库后,一次问答的简要过程如下:首先用户从终端输入问题,问句理解模块经过意图识别以及槽位提取和槽位改写等问句理解过程后,把得到的最终意图和槽位信息传递给问句转化模块。然后,问句转化模块进一步将用户意图和槽位转换为 sparql 查询语言。最后,答案检索模块借助领域知识库,完成 sparql 的查询,并返回简洁答案到用户终端上呈现给用户。

3 系统实现

3.1 知识库建立

对于知识库的建立,首先从豆瓣电影网站分析问答系统需要的信息。主要包括:电影基本信息(评分、剧情等等)、电影和人员的关系信息(导演、演员、编剧)、电影和地区、电影和类型以及电影和语言的相关信息。然后,选取 Protégé 作为知识库编辑本体的工具,构建知识图谱本体模型。图 2 所示为系统的知识图谱本体模型。

在构建了本体模型之后,选取 Scrapy 爬虫框架编写爬虫从豆瓣电影获取数据信息,并存入数据库。随后,按照定义的本体模型利用 D2RQ 技术把数据转换为资源描述框架(resource description framework, RDF)中的 N-Triples 三元组形式的数据,开放领域知识图谱 DBpedia 就是用这种格式来发布数据的。

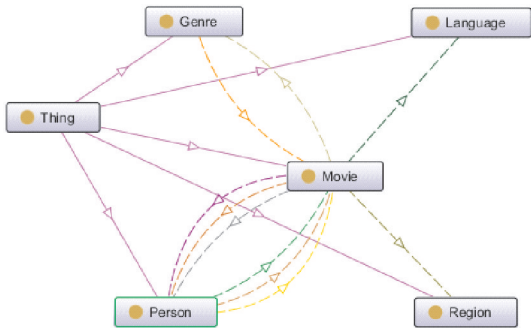


图 2 知识图谱本体模型

对于豆瓣的反爬虫策略,主要有两种解决方式:限速和爬虫代理。限速是通过控制爬虫速度的方式防止被反爬虫,爬虫代理则是通过多个代理 IP 进行多线程的信息爬取。爬虫代理分为免费代理和付费代理。免费代理可以从一些网站获取,但大多免费代理基本不可用。对于付费代理,可直接从某些网站购买,可用性虽强但实用性不高。文中系统采用的是限速方式,只要控制好速度,就可以越过豆瓣的反爬虫策略,目前的爬取速度为每 2 s 爬取一个页面。

由于电影数据会随着时间的不断增加和变化,因此爬取的电影信息需要更新,比如评分和评价人数等信息。文中系统的做法是:限速爬取豆瓣电影网站的信息,并定期增加和更新电影数据。截至目前,系统的知识图谱共包含 929050 个三元组。

3.2 意图识别

3.2.1 数据集

由于限定领域内的问题有限,几乎可罗列,因此在限定领域中问句意图识别的问题也就可以转换为一个分类问题。根据所构建的知识图谱本体模型,制定了 28 种意图。然后,通过模板构造问句(人工构造问句模板,借助领域词典填充问句模板获得问句)、百度知道以及朋友搜集的方式整理出了 2 万多条电影领域的问句数据集,作为问句分类的训练语料和测试语料。图 3 所示为部分数据集样本。

捉妖记2的评分 0
2017年上映的羞羞的铁拳怎么样,评分有多少 0
蒂姆·德莱克斯是什么星座 10
演员张立的外文名? 13
歌舞青春在制片国家有哪些 14
那部战争类型娘子军传奇的导演是谁 15
谁导演了电影我的早更女友 15
电影回到未来有哪些演员 16
理查德·尼克松演过的电影都有哪些风格 17
施正泉演的剧情电影 18
成龙和吴君如合作的电影有哪些 18
2014年上映的电影有哪些? 18
王默秋在2017年拍过哪些爱情类型的电影呢? 18
二零零九年的电影有哪些 18
有没有说上海话的电影? 18
07年上映的电影,评分多于八点6的有哪些? 19
电影遭遇大王奇遇记在哪一天上映的 1

图 3 用于分类的问题集样本

3.2.2 问句分类

采用深度学习中的 CNN 模型(卷积神经网络模型)用于问句的分类。经问句分析后发现问句都比较短,利用 jieba (python 实现的分词库) 切词,一个问句中的词语个数不超过 20。因此,将 CNN 嵌入层的输入设为 20 维的向量。问句一共 28 个类别,于是设置 CNN 模型的输出为 28 维。图 4 所示为最终的 CNN 模型结构。以下为问句分类的处理过程。

(1)通过切词后的问句集构建领域词典 Vocab,词典中每个词都有一个相应的数据序号。然后,利用该词典可以把一个切词后的问句转为一个数字向量。向量中的每个值代表对应的词在 Vocab 中的序号,长度不足 20 的左侧补 0。嵌入层采取随机初始化词向量的方式,设置每个词对应的向量为 100 维。于是,经过嵌入层之后,一个问句就转换成了 (None, 20, 100) 的一个张量。对于问句的所属类别,将其转换为一个 one-hot 向量,由于一共 28 个类别,因此 one-hot 向量为 28 维。

(2)将张量化后的问句及向量化后的所属类别向量送入 CNN 神经网络模型进行训练,以确定模型中的参数。测试时,系统先将用户输入的问题转换为 20 维向量输入到 CNN 模型,得到一个 28 维的 one-hot 向量输出。随后,将 one-hot 向量转为类别数字也就得到了问句的类别,即问句的意图。

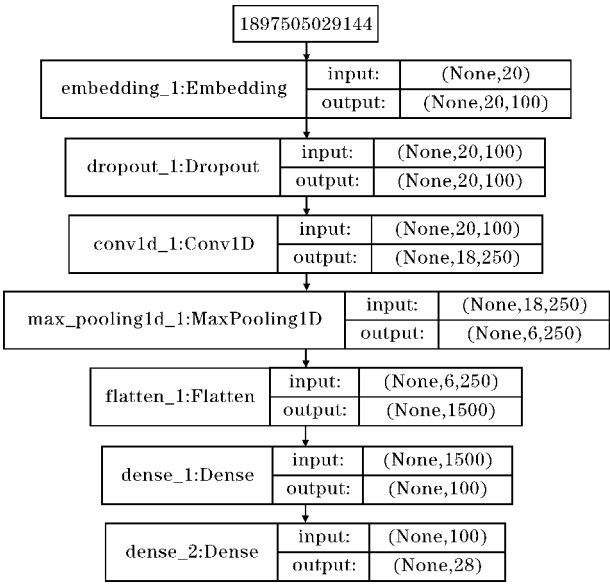


图 4 搭建的 CNN 模型结构

模型采用五折交叉验证的方式,将问句数据集进行 4:1 划分,4 份用于训练模型,剩下的 1 份用于验证测试,最终得到的平均分类准确率为 99%。

在该数据集上,也采用了其他的分类算法进行问题分类效果对比,实验结果如表 1 所示。

从对比结果可以看出,在该问句数据集上,基于关键字特征的 NB(朴素贝叶斯)分类算法效果最差,基于 CNN 模型的问句分类效果最好。

表 1 不同分类算法效果对比

算法	precision	recall	f1score
CNN	0.99	0.99	0.99
LSTM	0.97	0.97	0.97
MLP	0.90	0.90	0.90
LR	0.94	0.94	0.94
NB	0.85	0.85	0.85
SVM	0.96	0.96	0.96

3.3 槽位提取

在问句意图准确分类后,下一步是问句中槽位的提取,即问句中的关键信息。比如问句:“成龙一八年演过在美国地区上映过的动作电影有哪些?”,经过问句意图识别后,可以得知意图是电影查询。而槽位提取就是要提取出该问句中的关键信息,即:{演员:“成龙”;年份:“一八”;地区:“美国”;类型:“动作”}。再比如,问句“我不是药神的演员有哪些?”,需要准确提取出其中的电影名为“我不是药神”。

在知识图谱建立的同时,系统构建了电影名和人名词典,并将此词典加入 jiaba 分词利于人名和电影名的提取。但是,这种基于词匹配的实体识别方式不能做到理解问句的语意,可能把句子中任意位置匹配上的词语判定为人名或电影名。同时,这种方式无法识别类似“曼迪·摩尔”这种中文的外国人名。

此外,对于年份以及评分值槽位,如果直接采用词性标注提取槽位值可能会出现错误。比如,问句“长江 7 号的评分是多少?”,如果利用词性标注的结果:“长江/ns 7/m 号/m 的/uj 评分/n 是/v 多少/m”来提取槽位值,可能就把词性为量词的 7 作为了评分。究其原因,是基于词性标注或匹配的方式没有考虑问句的语意。

为此系统采用了序列标注模型 Bi-LSTM+CRF 结合词典进行槽位值的提取。词典用于提取地区、语言、电影类型等包含限定值的槽位,Bi-LSTM+CRF 用于提取人名、电影名、年份、以及评分值等槽位。

首先,对已有的问句集进行基于字的标注,作为 Bi-LSTM+CRF 模型的训练集和测试集,标注后的数据样本格式如图 5 所示。对于演员、年份等槽位只需要类似的定义标注即可,比如演员的标注定位为 B-PER 和 I-PER。

特别的爱的评分是多少

B-MV
I-MV
I-MV
I-MV
O
O
O
O
O
O
O

图 5 序列标注模型数据样本

得到标注的数据集后,下一步是搭建 Bi-LSTM+CRF 模型用以对数据集进行训练和测试。通过训练,Bi-LSTM+CRF 模型在本数据集上的槽位识别准确率为 98%。

模型训练结束后,即可用于对新问句的部分槽位提取。提取出的槽位再和使用字典提取出的其他槽位(如电影类型)结合,就可以获取出完整的槽位。

为了增加系统的实用性,在槽位提取完成后还需要进行槽位值的改写,系统采用同义字典进行槽位的改写。对于人名和电影名的同义词典,先从知识库获取人、电影的其他中文别名,进而构建别名词典。

陈港生一七年演过的评分大于八点七的电影有哪些?
{'pers': '陈港生', 'rate': '八点七', 'year': '一七'}

成龙2017年演过的评分大于8.7的电影有哪些?
{'pers': '成龙', 'rate': 8.7, 'year': '2017'}

图 6 槽位识别和改写效果

当输入问句“陈港生一七年演过的评分大于八点七的电影有哪些?”,经过字典和 Bi-LSTM+CRF 模型提取出的槽位,以及改写后的槽位和问句如图 6 所示。

3.4 问句转换和查询

在对问句完成意图分类和槽位提取后,下一步是将其转换为知识库可识别的查询语句。对于限定领域内的有限意图可先制定相应的查询语句模板,然后将提取并改写完成的槽位信息填入查询语句模板中,就可以生成相应的 SPARQL 查询语句。

对于问题“陈港生一七年演过的评分大于八点七的电影有哪些?”,生成的查询语句如图 7 所示。

```
PREFIX : <http://www.auto_answer_for_movie.com#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?o
WHERE {
    ?m :movieTitle ?o .
    ?m :movieYear '2017' .
    ?s1 :personName '成龙' .
    ?m :hasActor ?s1 .
    ?m :movieRate ?a .
    Filter(?a < 8.7) .
}
```

图 7 SPARQL 查询语句示例

系统采用 Fuseki 作为 SPARQL 服务器,它可以通过 HTTP 提供 RDF 数据。配置好 Fuseki Server 之后,上传构建的电影领域 RDF 三元组数据,知识库检索服

务就搭建完成。随后,问句检索模块通过 HTTP 请求发送生成的 SPARQL 语句到 Fuseki Server 进行查询,即可得到相应的结果。

3.5 系统运行效果

为了便于用户使用,系统采用微信公众号作为客户终端。

首先,以开发者的身份注册一个微信公众号。然后,将基于 Django 框架的电影问答系统作为公众号的后台问答服务。最终运行效果如图 8 所示。

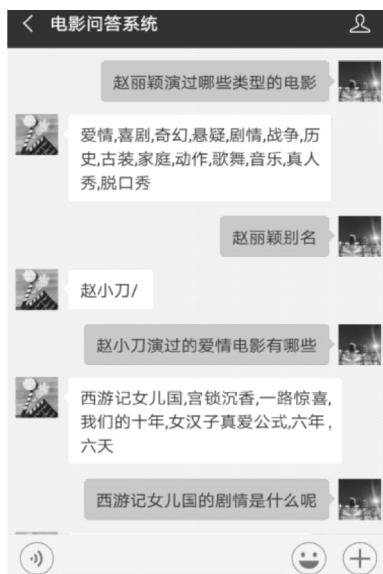


图8 问答系统效果图

最后,为了系统后续的不断优化,比如对暂时无法回答的问题,以及错误分析问题进行原因分析。系统增加了日志模块,记录所有用户提过的问题、系统的意图分析和槽位提取结果。

4 结束语

针对现有的领域问答系统发展不够完善的现状,提出了将深度学习算法应用于问答系统中,并借助深度学习算法,设计并实现了一个基于知识图谱的电影领域问答系统。系统的各模块耦合性低,具有很好的扩展能力,可以比较方便地增加意图或者向其他领域扩展。

目前,系统可以对常见的领域问题进行回答。但由于制定的意图有限,因此超出范围的问题系统会答复“暂时无法回答”。同时,基于搜集和模板构建问句数据集的方式只能覆盖大部分用户的问句表达方式,无法完全覆盖用户表达。因此,基于目前数据集训练的意图识别模型,对某些未训练到的问句表达方式会

出现无法识别的情况。此外,目前系统只是针对中文问句的处理,因此知识库也只获取了中文名电影和演员的信息,未获取英文电影相关信息。

未来可考虑直接利用深度学习算法深层理解问句,并端到端地直接得到查询语句,进而避免手工制定意图类别及查询语句模板的繁琐工作。

参考文献:

- [1] 夏元昉. 基于深度学习的问答系统技术研究[D]. 杭州:浙江大学,2017.
- [2] 邢世祥. 基于深度学习的智能问答系统研究[D]. 北京:北京邮电大学,2016.
- [3] Erfan Najmi, Khayyam Hashmi. Intelligent semantic question answering system[C]. IEEE International Conference on Cybernetics, 2013:255-260.
- [4] 杨志明,王来奇,王泳. 深度学习算法在问句意图分类中的应用研究[EB/OL]. <http://kns.cnki.net/kcms/detail/11.2127.TP.20180913.0626.002.html>, 2018-09-14.
- [5] Green B, Wolf A, Chomsky C, et al. BASEBALL: an automatic question answerer[M]. Readings in natural language processing, Morgan Kaufmann Publishers Inc, 1986:545-549.
- [6] Weizenbaum J. ELIZA. A computer program for the study of natural language communication between man and machine[J]. Communications of the Acm, 1966, 9(1):36-45.
- [7] 詹晨迪. 基于知识库的自然语言问答方法研究[D]. 北京:中国科学技术大学,2017.
- [8] 王东升. 面向限定领域问答系统的自然语言理解方法综述[J]. 计算机科学, 2017, 44(08):1-8.
- [9] Spink A, Gunar O. E-commerce Web queries: Excite and Ask Jeeves study[J]. 2001, 6(7).
- [10] Katz B, Borchardt G C, Felshin S. Natural Language Annotations for Question Answering[C]. FLAIRS Conference. 2006:303-306.
- [11] 陶永芹. 专业领域智能问答系统设计与实现[J]. 计算机应用与软件, 2018(5):95-101.
- [12] 耿新鹏. 基于深度学习知识库问答研究进展[EB/OL]. <http://blog.openkg.cn/技术动态-基于深度学习知识库问答研究进展/>, 2018-11-25.
- [13] 李超. 基于深度学习的短文本分类及信息抽取研究[D]. 郑州:郑州大学,2017.

Design and Implementation of Domain Question Answering System based on Deep Learning

HU Jie, TAO Hongcai

(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Intelligent question answering system can provide users with information services quickly and accurately, which is a research direction that attracts much attention in the field of natural language processing. This paper aims to implement a domain question answering system based on knowledge base with deep learning algorithm. In the construction of knowledge base, firstly, Protégé editing ontology model is used. After completing domain knowledge acquisition and RDF knowledge graph construction, Fuseki Server is used as SPARQL server to provide query interface for the query module of question answering system. In question comprehension, by comparative experiments of a variety of classification algorithms, CNN model of deep learning is selected for question classification. Then, the sequence annotation model Bi-LSTM+CRF is flexibly applied to slot extraction in question sentences. Finally, a question-and-answer system in the field of film based on WeChat Subscription is completed. The test results show that in the question set, the accuracy of question comprehension can reach 98%, and based on the domain knowledge map, most domain questions can be answered more accurately and quickly.

Keywords: question answering system; deep learning; knowledge graph; question classification