面向社交群问答对获取的深度学习模型

张民航 蔡东风 李绍鸣
(沈阳航空航天大学人机智能研究中心 辽宁 沈阳 110135)

摘 要 关注社交群中的问答资源,提出面向社交群的问答对获取方法,主要包括问句识别和答案获取。分析了基于规则和深度学习及结合方法三种问句识别方法的特性;答案获取以深度学习模型为基础,将区分正反例回答同问题的相关度作为学习目标,对各个候选答案与问题的相关度打分排序。引入回答顺序和共现词特征对基础打分作调整进行二次打分排序。实验结果表明,问句识别方法在 WebQA、Dbqa 和真实小区群聊语料 CMY 上的 F1 值分别达到 0.930、0.932 和 0.892;CMY 上的问答对获取 F1 值达到了 0.690。

关键词 问答对获取 问句识别 问答匹配 问答系统

中图分类号 TP391.1 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.028

DEEP LEARNING MODEL FOR ACQUISITION OF Q&A PAIRS IN SOCIAL GROUPS

Zhang Minhang Cai Dongfeng Li Shaoming
(*Human-Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang 110135, Liaoning, China*)

Abstract Focusing on QA resources in social groups, this paper proposes a QA pair acquisition method oriented to social groups, which includes question recognition and answer acquisition. The characteristics of three different methods including rule-based method, deep learning based method and combined method were analyzed. Answer acquisition was based on a deep learning model, and the correlation between positive and negative example answers and the question was taken as the learning objective, and the correlation between each candidate answer and the question was scored and sorted. The answer order and word co-occurrence features were introduced to adjust the basic scores for secondary scoring and ranking. Experimental results show that the F1 scores of the question recognition method are 0.930, 0.932 and 0.892 on WebQA, Dbqa and real community group chatting corpus CMY respectively; the QA pair on CMY achieves an F1 score of 0.690.

Keywords QA pair acquisition Question recognition Q&A matching Q&A system

0 引 言

近几年关于中文问答系统的研究日渐火热,也暴露出了一些问题,如问答语料匮乏、人工构建语料困难、高质量问答领域单一或广域问答质量不高等。实际上,网络上存在大量的问答资源,若能好好利用这些资源,可为问答系统的研究提供丰富的语料资源。当前有关问答对抽取的研究如文献[1-4]都集中于网页资源,如论坛、社区、贴吧和网站 FAQ 等。除了上述网页资源,网络上还存在其他问答资源,比如各类社交群资源。随着各类社交群如微信群、QQ 群和微博群的兴起,其中积累了大量的问答资源,但是这些资源尚未得到研究者的重视。面对这一状况本文开展了面向社交群的问答对获取研究。

社交群资源不同于网页资源,后者多是一问一答或一问多答,社交群成员众多、发言自由,大都是多问多答。本文着重研究了问句识别、回答顺序、候选答案范围等影响问答对获取质量的关键因素,提出了通用的问句识别方法和新的问答对获取方法。该问句识别

收稿日期:2020-01-06。教育部人文社会科学研究项目(17YJCZH003);辽宁省重点研发计划项目(2019JH2/10100020)。
张民航,硕士生,主研领域:人工智能,问答系统。**蔡东风**,教授。**李绍鸣**,硕士生。

方法不仅可用于问答对获取,也能在问答系统中过滤非问题干扰;所提问答对获取方法基于问答匹配策略实现,亦能用于问答系统的答案匹配过程。

1　相关工作

问答对获取,即从含有问答内容的文本中自动地获取问题和答案。通常,问答对获取分为两步完成:问题识别(又名问句识别)和答案识别。

目前,关于中文问句识别的研究可以大致分为两个类型:(1) 从语言学的角度识别问句,该类方法大多借助特征词、句式、句法结构和语法单位等特征判别问句。(2) 从文本分类的角度识别问句,这类方法大多借助机器学习,通过有监督训练的方式将问句识别视作二分类问题。周飞云[5]总结了设问句和反问句的类型及特征,并分析了设问句与反问句嵌套的特殊情况。殷树林[6]提出了反问句的三个基本特征:无疑而问、不需要回答和表示否定,进一步细化了反问句的句法结构,总结了 17 类反问句特有的句法结构。与上述研究从语言学角度分析反问句特征从而人工判断反问句不同,文治等[7]分析了反问句的句式特点,将反问句的句式结构融入到卷积神级网络的构建中,实现了中文反问句的自动识别。其在微博反问句上的识别准确率、召回率和 F1 值分别达到了 89.5%、84.2% 和 86.7%。熊作平[8]分析了汉语是非问句的类型,并从语法层面归纳了 3 种肯定是非句格式和 7 种否定是非句格式。陈彩霞[9]着重分析了标点符号、语气词、语调等特征在是非句识别中的作用。侯永帅等[10]针对问答系统中部分问句对时效敏感的特性,设计了时间敏感问句的识别和检索方法。实验表明该方法是有效的,使用 C5.0 决策树识别时间敏感问句准确率可达 0.901。

在识别问句的前提下进一步识别答案才能获得完整的问答对。文献[1-2]根据 HTML 内容建立 DOM 树,将树中各个节点识别为问题(q)、其他(o)和答案(a)三类,取树中以 QA 顺序出现且距离最近的一对为问答对。文献[3-4]研究社区问答对获取,从用户回帖中识别答案,通过对回帖分段划分再使用文本特征和非文本特征训练机器学习模型,进而预测答案概率。当前,越来越多的研究者开始使用深度学习算法开展问答匹配研究,这类方法不需要太多的人为干预,可以自动学习问答匹配特征并加以利用。此外,词向量和注意力机制的引入进一步加强了深度学习处理问答匹配的能力。使用深度学习进行问答匹配需解决两个关键问题[11]:(1) 实现自然语言问句及答案的语义表示;(2) 实现问句及答案间的语义匹配。荣光辉等[12]面向中文问答匹配任务,提出了基于组合模型的问答匹配方法。邢超[13]通过问句向量生成模型和答案向量生成模型分别将问句和答案表示为特征向量,同时将两个模型组合成混合向量模型用于问答匹配。Shen 等[14]针对词袋模型难以在短文本匹配中捕获重要的词序列信息这一问题,提出了一种新的体系架构。Yang 等[15]将问答对匹配视为二分类任务,并提出了对抗性训练框架来减轻标签不平衡问题。Wu 等[16]提出了 FACM 模型,该模型通过使用匹配策略扩展卷积神经网络(CNN)来对涉及多方面主题的 QAP(问答对)进行注释。Lima 等[17]提出了一种基于知识的问答框架,该框架将多级标签推荐与外部知识库集成在一起,以检索最相关的知识库文章来回答用户发布的问题。

综上所述,当前的问句识别研究主要集中于特殊问句的识别,并没有适用于一般问句的识别方法,而社交群中的问句类型丰富不仅包含特殊问句。因此,本文提出了通用的问句识别方法。鉴于目前问答对获取的相关工作并未考虑候选答案范围和答案顺序,本文分析了候选答案范围对问答对获取质量的影响,提出了结合答案顺序的问答匹配框架。

2　社交群问答对特性分析

面向社交群的问答对获取,即从源于社交群的文本(一般是群聊记录)中成对地找到问题及其答案。为此,需要解决两个关键问题:问句识别;候选答案范围的确定。各类社交群的文本,如微信群文本、QQ 群文本、钉钉群文本等,以下简称群记录。

2.1　问句特点

问句和非问句的区别主要体现在词语、句式、符号、句尾和短语搭配等方面,此外,出于易于理解的目的,问句一般都比较简短。本文归纳了问句的一般特征,在此结合统计结果予以说明。

由经验不难发现,某些词出现在问句中的比例远高于在非问句中的比例,如各类疑问词“谁、哪里、何时、为什么、怎么样”等。问句有自己特有的句式结构:(1) 一般疑问句,疑问词多出现在句子首尾,[疑问词,句中,句尾]或[句首,句中,疑问词]。(2) 反问句,句法结构较多可参考文献[6]。(3) 选择疑问句,

是 A 还是 B(A、B 均为字符串)。(4) 是否疑问句,...CDC...结构(C 为词语 D 为否定字或词),如“电影票还有没有”、“这衣服好看不好看”。问句的符号特征主要是“?”,句尾特征主要有“吗”、“么”、“没”、“不”等。实际上句尾特征“没”和“不”属于句式特征(2)的省略用法,如“你吃饭没吃饭?”,省略后为“你吃饭没?”;“你买不买某物?”,省略后为“你买不?”。

本文将中文问句分为两大类:显性问句和隐性问句,其中显性问句又分为强制问句和一般问句两小类。显性问句指含有显性特征的问句,隐性问句指不包含显性特征的问句。

显性特征主要有符号特征和句尾特征,除错误使用符号和句尾字的情况外,一般含有此类特征的句子皆可视为问句。在真实的交流情景中,人们可能会通过语气而非泛指信息表达疑问。比如“我要是不去,你来。”表达的是疑问语气“难道你来?”,这类情况通过语气强行使之成为问句,若不通过语气(语调)仅从字面上则无法判断是否为问句。

与上述口语现象对应,书面语也有类似现象,即通过“?”使句子强行成为问句,一旦去除问号则难以判断其为问句。比如“大米 5 元 1 斤?”是询问大米的价格,而“大米 5 元 1 斤”通常都不会被视为问句。同时,这类情况也难以使人判断此处是否存在符号误用,也许他人正是想通过此类方式表达疑问。这种通过问号强行变成问句的句子即为强制问句。而一般问句大多含有疑问词或疑问句句式特征,即使去除问号,也基本可以判定为问句,如“大米多少钱一斤?”“你去不去买大米?”。

显性问句可通过显性特征来判断,问句识别的主要工作也由此转为隐性问句识别。

2.2 答案特点

候选答案范围(即用于获取答案的文本片段)的确定对正确获取问答对有重要意义,其决定了是否可以获取答案以及获取答案的效率。为便于说明,引入三个定义:(1) 话语角色,即问答的角色,主要有提问者和回答者。(2) 角色职能,提问者提出问题,回答者回答问题。(3) 答案窗口,即候选答案的范围。群记录中话语角色的数量即是群成员的数量,每个成员既可能是提问者也可能是回答者。群记录中的问答存在以下几个特点:(1) 答案不一定存在,问答皆是自由的,某些问题无人回答。(2) 答案源于问题之后的若干人次发言,其与群规模无必然联系,与活跃人数关联较大。(3) 问题之后的发言不一定都是回答,很可能有其他问题。(4) 很多问题都是对某些信息进行确定,因此答案多是短文本。(5) 同一问题,可能有多个相同或不同的回答,应获取最适合的回答。

本文在确定答案窗口时,依据两个假设:答案范围不超过问题之后最近的 10 人次发言;不能以问题来回答问题,即答案不会出现在问句中。通常,答案窗口越大越可能获得更多的问答对;同时,越大的答案窗口包含了越多的干扰,获取问答对的质量有可能降低。

3 问句识别

3.1 方法介绍

3.1.1 基于规则的问句识别方法

问句和非问句在句式、用词、符号、短语搭配等方面有明显的区别,本文总结了中文问句的一般特征,由此提出了基于规则的问句识别方法 RBQR(Rule Based Question Recognition)。实际上虽然句尾特征和“?”存在误用的情况,但其比例非常小,因此本文直接将句尾含有显性特征的句子视为问句,所提出的基于规则的问句识别方法也主要用于识别隐性问句。本文所用规则主要包括三部分:显性特征、“是”规则与“否”规则。识别问句的基本思路是:(1) 含有显性特征“吗”“?”“么”“不”“没”的句子直接被判定为问句。(2) 通过“是”规则集初步判断候选问句集。(3) 使用“否”规则集过滤候选问句集,得到最终问句集。“是”规则主要包括八大特征:符号、句尾词、特征词、句法特征、固定短语、用语习惯、句长、特征词词序。“否”规则主要包括固定搭配和否定特征词两类特征,如,特征词 + <都,也>,否定动词 + 特征词。

符号特征,即句尾是否含有问号。

句尾词,主要有 <吗,么,没,不>。

特征词,主要有 <什么,谁,哪,怎么,多少,几,咋,哪个,哪里,请问,谁家,怎么样,咋样,为什么,如何,什么样,多久,怎么办,哪些,哪家,哪位,多大,多长,多宽,多高,多远,多重,多快>。

句法特征,主要包括特征词的句法标记和其下一级依赖词的句法标记;词序,主要是句首和句尾特征。句法特征与词序特征配合使用,大体分为四种情况,具体关系见表 1。

表 1　句法特征与词序特征搭配的规则

特征词	特征词标记	依赖词标记	句首距离	句尾距离
为什么	HED ADV ADV VOB	\ HED COO \	\ \ \ \	\ \ \ <3
谁,谁家,哪位,哪里,何时,何地,何处,何年,何月,多久	SBV VOB	\ \	<3 \	\ <3
什么,哪,哪个,哪些,多快,多少,几,多大,多长,哪家,多宽,多高,多远,多重	HED SBV VOB ATT ATT ATT ATT	\ \ \ HED ADV SBV VOB	\ <3 \ \ \ <4 \	\ \ <3 \ \ \ <4
怎么,咋,怎么样,咋样,怎么办,如何	HED ADV SBV VOB	\ HED \ \	\ \ <2 \	\ \ \ <3

表 1 说明如下:(1) 特征词,用于问句识别的关键词。(2) 依赖词,特征词在句法分析中的后继。(3) 标记,即词的句法标记,表示当前词与其所指词的句法关系。其中,HED 为核心关系;SBV 为主谓关系;VOB 为动宾关系;ADV 为状中结构;ATT 为定中关系;COO 为并列关系。(4) 句首距离,表示特征词在当前子句(以"?!,;。"对长句切分所得即为子句)中的顺序词序。如"还有 谁 没来 ?","谁"的句首距离为 2。(5) 句尾距离,表示特征词在当前子句中的倒序词序。如"你 喜欢 吃 什么 水果 ?","什么"的句尾距离为 3。(6) "\"表示该项为空。规则使用示例如图 1 所示。

示例:

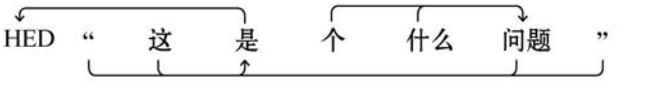

图 1　规则使用示例

图 1 中上半部为分词后的待判断句子,下半部为各词对应的[依赖词位置:当前词句法标记]。特征词是"什么",其句法标记"ATT",依赖词为"问题",依赖词属于宾语成分句法标记"VOB",特征词句尾距离为 2。因此,此句按规则判断为问句。

固定短语,仅凭特征词还无法直接判断一个句子是否为问句,但是当特征词与某些词形成搭配短语时,基本可以断定一个句子为问句。比如 < 谁有,谁知道,谁的,干嘛,是什么,有什么,想问一下 > 。

我国地域辽阔,各地的人用语习惯不同,这些口语上的习惯也会反映在文字上。比如东北地区的特有用语习惯 < 干啥,是不,吃饭没,出去没 > ;四川省的特有用语习惯 < 干啥子,你晓得…,哪个(谁) > 。这些用语习惯或为句式特征或为短语特征,可作为重要的问句识别特征。在对一些来自局部地区的文本特别是口语化较重的文本进行问句识别时,用语习惯特征会有很大作用。

句长特征,为使被问者清楚地了解自己的意图,提问者通常会用较为简洁的语言描述问题,这也使得问句文本一般较短。可以通过对显性的问句统计,适当估计问句的长度,以此筛选掉过长的句子,从而提升问句识别的精度。

基于规则的问句识别流程如图 2 所示。

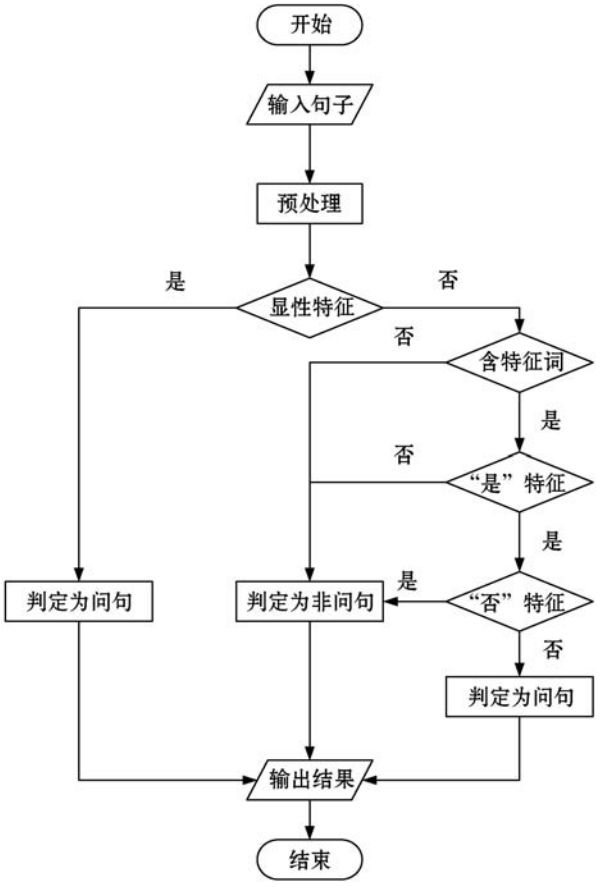

图 2　基于规则的问句识别流程

对于任意新输入的句子,先对句子进行分词、词性标注和句法分析等预处理,随后判断句子中是否含有显性特征,含有显性特征则判定为问句,否则进入后续流程。在后续处理过程中将不含特征词的句子直接过滤掉,反之判断是否包含"是"特征,再接着判断是否含有"否"特征。经过多重处理,最后得到的句子就判定为问句。

3.1.2　基于深度学习的问句识别方法

基于规则的问句识别方法可以借助规则准确地识别问句,但是规则由人工制定扩展亦需耗费人力,且规则不易覆盖所有问句类型。为了从不同角度探索问句

识别的特性和效果,本文亦使用深度学习的方法开展问句识别研究,主要包括基于 CNN 的问句识别方法和基于 LSTM 的问句识别方法。

目前,CNN 网络和 LSTM 网络已被广泛应用于自然语言处理领域且表现出色,前者具备较强的局部特征学习能力,后者具备较好的整体语义把握能力。本文将问句识别视为二分类任务,使用有监督学习的方式,分别训练 CNN 模型和 LSTM 模型预测待识别句子,以 0.5 为阈值将大于此值的句子判定为问句,反之为非问句。

3.1.3 规则与深度学习相结合的问句识别方法

基于规则的问句识别方法和基于深度学习的问句识别方法各有所长,本文将两者结合在一起用于问句识别。该结合方式并非将两种方法融合为第三种方法再进行问句识别,而是由两类方法分别预测出候选问句集,取两个候选集的并集作为最终预测结果,其结果构成见图 3。

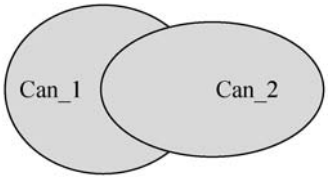

图 3 结合方法问句识别结果构成图

具体结合方式如下:(1) 由 RBQR 方法预测出候选问句集 Can_1,即图 3 左半部;(2) 由 CNN 模型或 LSTM 模型预测出候选问句集 Can_2,即图 3 右半部;(3) 对 Can_1 和 Can_2 取并集 Can,即图 3 整体,以此作为最终识别结果。

3.2 实验与结果分析

3.2.1 数据集和评价指标

实验所用数据集共包含三部分:小区群聊语料 CMY,百度公开的开放域问答数据 WebQA[18] 和 NLPCC-ICCPOL2016 开放域问答数据 Dbqa。其中,CMY 为本文收集整理的真实语料,源自某小区微信群聊记录,主要包含了用户的提问和相关的回答,夹杂了少量闲聊内容。为保护隐私所有数据均进行了脱敏处理。此三种语料皆是开放域问答数据,所含的问题类型丰富、涉及领域广泛,能够很好地体现问句特点代表性强。

CMY、WebQA 和 Dbqa 的基本构成如表 2 所示。

表 2 问句识别数据集简介 单位:个

数据集	隐性问句	显性问句	非问句
CMY	1 481	1 429	12 120
Web_data	2 000	\	16 000
Db_data	2 000	\	16 000

问句识别实验分别在 CMY、Web_data 和 Db_data 三个数据集上进行,采用查准率 P(Precision)、召回率 R(recall)和 F 值(F-Measure)作为评价指标。

3.2.2 实验设置

问句识别实验分别使用基于规则的方法、基于 CNN 模型和基于 LSTM 模型的方法在 CMY、Web_data 和 Db_data 数据集上进行,共计 9 组实验。其中,基于规则的方法采用本文提出的中文问句识别方法 RBQR。

CNN 模型基本设置:(1) 输入层,采用嵌入层,词向量维度 128。(2) 卷积层,256 个卷积核,卷积核大小为 10,激活函数为 ReLU。(3) 池化层,采用最大池化方式。(4) 输出层,采用全连接层,维度为 1(二分类,输出层即是 1 维),激活函数为 sigmoid。(5) 损失函数 loss = 'binary_crossentropy',优化器 optimizer = 'Adam'。

LSTM 模型基本设置:(1) 掩蔽层,timesteps = 50。(2) 输入层,采用嵌入层,词向量维度 128。(3) LSTM 层,维度 64,dropout = 0.2,recurrent_dropout = 0.2。(4) 输出层,采用全连接层,维度为 1(二分类,输出层即是 1 维),激活函数为 sigmoid。(5) 损失函数 loss = 'binary_crossentropy',优化器 optimizer = 'Adam'。

3.2.3 结果分析

问句识别结果如表 3 所示。

表 3 问句识别结果

Dataset	Method	p	R	F1
CMY	RBQR	0.911	0.728	0.809
	LSTM	0.938	0.747	0.832
	CNN	**0.943**	0.776	0.851
	RBQR + LSTM	0.881	0.883	0.882
	RBQR + CNN	0.885	**0.900**	**0.892**
Web_data	RBQR	0.927	0.882	0.904
	LSTM	**0.975**	0.885	0.928
	CNN	0.974	0.880	0.925
	RBQR + LSTM	0.866	0.935	0.899
	RBQR + CNN	0.907	**0.955**	0.930
Db_data	RBQR	0.918	0.900	0.909
	LSTM	0.908	0.927	0.917
	CNN	**0.928**	0.936	**0.932**
	RBQR + LSTM	0.836	**0.983**	0.904
	RBQR + CNN	0.859	0.978	0.915

从表 3 中可以得到:

(1) 选择 Web_QA 和 Dbqa 数据与 CMY 对比以及对应的数据抽取策略是合理的,三种问句识别方法在各个数据集上的表现基本一致:识别准确率均大于 0.9,

CNN 识别效果与 LSTM 识别效果接近且 CNN 方法略强于后者。

(2) 基于规则的方法 RBQR 在问句识别任务上识别准确率较高,召回率较低。主要原因有:RBQR 针对问句特性设计规则,可以充分利用问句特征;RBQR 判别标准较为严格,部分不符合标准即可能被判定为非问句;规则很难完全覆盖各类情况,一些不符合规则却是问句的句子难以识别。

(3) 将问句识别视作文本分类任务是可行的,在所测试数据集上基于深度学习的方法 CNN 和 LSTM 均有良好表现,其中 CNN 方法基本在各个数据集上达到了最佳识别效果。主要原因有:LSTM 模型能很好地从整体上把握问句,问句大都是短文本,大部分在 30 个词以内,LSTM 的记忆特性在此可以得到较好的发挥;问句的特征较为明显,如疑问词、固定短语等特征,CNN 可以轻松地卷积出此类特征,所以 CNN 模型在问句识别任务中表现很好;相比句子的整体特性,问句识别对局部特征更为敏感,这使得 CNN 模型比 LSTM 模型表现更好。

(4) 基于规则的问句识别方法与基于深度学习的问句识别方法具有较好的互补性,其结合方法召回率最高。当单一方法召回率较低时,其能显著提高召回率且可能因此提升 F1 值,CMY 和 Web_data 的最大 F1 值皆源于结合方法可体现此特性。随着召回率的提升,融合方法的提升效果也随之减弱。

整体而言,问句特征和语义都在问句识别中起重要作用。RBQR 可以针对问对特性设计规则具备较高的"判正性",但是规则较为严格且对非问句没有针对性处理;CNN 和 RNN 可以学到问句和非问句特征,能在具备较高识别准确率的同时具备较高召回率。将二者结合起来,可在适当牺牲准确率的情况下提升召回率,在召回率较低时提升效果尤为显著。

识别错误的问句大体可分为两种:(1) 强制问句去除问号后不含任何问句特征,几乎不可识别。如 CMY 中"你买房子?"变成"你买房子",则无法识别为问句。(2) 非强制问句疑问词省略,去除问号后同问题(1)。如 Web_data 中"爱因斯坦的老婆是(谁)?"变成"爱因斯坦的老婆是","谁"被省略则无法识别为问句。

4　问答对获取

4.1　基于深度学习的问答对获取方法

将问题与答案进行匹配的过程即为问答匹配,本文通过基于深度学习的问答匹配方法来获取问答对。当前关于问答匹配的研究大多是利用词向量将问题与答案映射到语义空间,再通过深度学习模型学习两者间的匹配特征,进而判断两者是否相符,这类方法将问答匹配视为分类问题;或通过深度学习模型计算问题与答案的匹配度,对问题与多个候选答案的匹配度打分排序,将其作为排序问题处理。由于一个问题可能对应多个候选答案,仅以问答是否相符为衡量标准是无法区分复数个候选答案优劣的,因此本文将问答匹配视为排序问题处理,将排在第一位的候选答案作为匹配结果。先通过深度学习模型对问题的各个候选答案打分排序,以此作为基本排序;再引入候选答案顺序、共现词等特征对候选答案二次打分排序。

通过双向 LSTM 实现孪生神经网络模型 TwM,以此学习问题和答案的深层特征。该模型的主要学习目标是区分错误答案和正确答案与问题之间的距离,即使得正确答案与问题的联系更紧密,而错误答案与问题的联系更疏远。实际上,问答匹配多数是从含有一个答案句子多个非答案句子的段落中获取答案,为了使模型较充分地学习到正例与负例的区别,通常应是一个问题对应一个正例和多个负例。

TwM 的网络结构如图 4 所示。

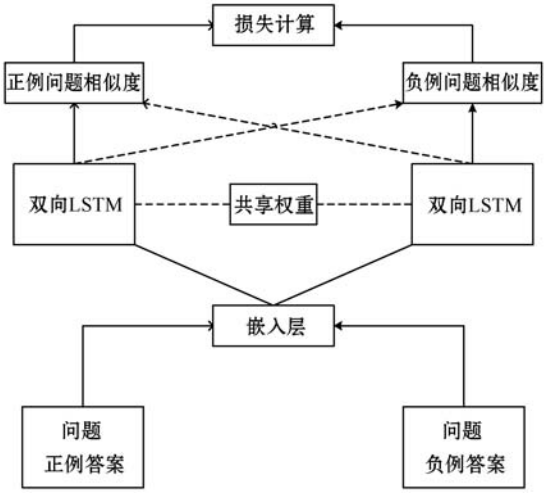

图 4　TwM 网络结构

采用孪生网络结构,共包含五层,自底向上依次是输入层、嵌入层、双向 LSTM 层、问答相似度计算层和损失计算层。该孪生网络一部分用于生成问题的特征向量,另一部分用于生成正反例的特征向量。图 4 中空心箭头代表问题和正例答案的工作流,实心箭头代表问题和负例答案的工作流。

输入层每次接受两个输入,一个是问句另一个是候选答案句子(正例或负例)。问题和正例是重复使用的,以便与负例数目配平。如 1 个问题、1 个正例和 3 个负例时,{[(q,Ta),(q,Fa1)];[(q,Ta),(q,Fa2)];[(q,Ta),(q,Fa3)]}为一个完整样本,q 表示

问题,Ta 表示正例,Fa1、Fa2、Fa3 依次对应三个负例。

嵌入层负责将输入层的句子(问句或正反例)表征为句子向量,以便映射到语义空间。句子向量由句子中的词对应的词向量相加后除以词数得到,词向量使用了基于中文维基百科预训练的词向量。

两个双向 LSTM 层通过共享权重实现权重一致,以保证对问句和正(反)例进行同样的特征提取。问题和正(反)例通过该层后将被表示为两个特征向量,进入下一层计算相似度。

问答相似度计算层也是实质上的输出层,实际使用模型时所需的即是该层计算的相似度。以上一层输出的两个特征向量的余弦距离来衡量问题与正(反)例的相似度,并以此作为 TwM 对候选答案的打分,记为 D_score,$0 \leqslant D\_score \leqslant 1$,其计算方式如下:

$$D\_score = \cos(\boldsymbol{a}, \boldsymbol{q}) \tag{1}$$

式中:$\boldsymbol{a}$ 是候选答案经 LSTM 层输出的向量,$\boldsymbol{q}$ 是问题经 LSTM 层输出的向量。

最后一层是损失计算层,主要用于计算损失函数。与一般有监督训练模型不同,TwM 属于无监督训练,并无标签可用,只能在每次计算出问题正例相似度和问题负例相似度后再计算损失。损失函数 loss 定义如下:

$$loss = \max\{0, margin - (t\_sim - f\_sim)\} \tag{2}$$

式中:$margin$ 为设置的阈值,t_sim 为问题正例相似度,f_sim 为问题负例相似度。$loss$ 的计算分为三步:(1)计算问题正例相似度与问题负例相似度的差。(2)计算 $margin$ 与步骤(1)结果的差。(3)取 0 和步骤(2)结果这二者中最大的作为损失。

以 D_score 的打分直接排序作为基本排序,记为 Rank。在此基础上结合候选答案顺序进行二次打分排序,记为 Re_rank,二次打分得分记为 Re_score。基本原则有二:(1)候选答案的次序距离问题越近,其成为答案的可能性越高。(2)候选答案的次序距离问题越近,其在二次打分中得到的增幅越大。

Re_score 的计算方式如下:

$$Re\_score = D\_score \times (1 + \lambda)^{\mathrm{ord}} \tag{3}$$

此计算方式对各个候选答案的增幅随次序呈指数级变化。式中:λ(一般取 0 ~ 0.1)为基本增幅,ord 为窗口中各候选答案到窗口底端的次序。如窗口为 4 时,有例子 1:Rank = [0.5,0.4,0.6,0.3],取 $\lambda = 0.1$,则 Re_rank = $[0.5 \times 1.1^3, 0.4 \times 1.1^2, 0.6 \times 1.1^1, 0.3 \times 1]$。

为便于表述,引入一些说明:以 Rank 排序为基准,若答案在候选答案中得分最高标记为 g,否则标记为 m,其余候选答案标记为 n。二次打分排序的目的就是使得排序靠前的候选答案 m 有机会借助顺序优势变成 g,从而获得更多的正确问答对。在例 1 中,若 Rank 标记为[m,n,n,n]即第一候选答案是答案,则 Re_rank 标记为[g,n,n,n]。在基本排序中答案的得分 0.5 < 0.6 不是最大的,而在二次打分排序中,答案的得分成了所有候选答案中最高的,因此可与问题配对为正确的问答对。

Re_rank 对顺序靠前的候选答案有所偏好,在二次排序时会导致原本标记为 g,但顺序靠后的答案失去高分优势,从而导致答案得分不再是最高的以致问答对获取失败。如例 1 中,若 Rank 标记为[n,n,g,n]即第三候选答案是答案,则 Re_rank 标记为[n,n,m,n],这将导致原本可以正确获得的答案因排序靠后而获取失败。以本文所用数据 CMY 为例:窗口为 4 时 Rank 的排序结果中标记“g”422 个,标记“m”0 个(排序结果取各组得分最高者,标记 m 非最高,不会被选中);Re_rank 的排序结果中标记“g”382 个,标记“m”52 个(二次排序后,标记 m 可能变成最高分被选中)。在二次打分排序后正确获取的问答对共 434 个,较 Rank 结果 422 多出 12 个,其中标记“m”的问答对增加了 52 个,但是标记为“g”的问答度丢失了 40 个。为了减少在二次排序中损失的“g”问答对,本文引入了共现词特征。统计结果表明,大多数标记为“g”的候选答案与问题有公共词语(即共现词),而标记为“m”的候选答案大都不含共现词。结合共现词的二次打分排序记为 Re_rank_com,其得分记为 C_score,基本原则是维持 Rank 中有共现词的“g”标记候选答案,具体计算方式如下:

$$C\_score = D\_score \times c + (1 - c) \times Re\_score \tag{4}$$

$$c = \begin{cases} 1 & \text{rank 中 g 标记候选答案有共现词} \\ 0 & \text{rank 中 g 标记候选答案无共现词} \end{cases} \tag{5}$$

式中:c 为共现词标记。

以 CMY 为例,结合了共现词的二次打分排序中标记“g”415 个,标记“m”30 个,正确获取的问答对共计 445 个较 Rank 排序 422 多出 23 个。可见,引入共现词化解 Rank 和 Re_rank 的冲突还是有效的。

4.2 数据集和评价指标

问答对获取实验在 CMY 数据集上进行,使用准确率查准率 P(Precision)、召回率 R(recall)和 F1 值(F1-Measure)作为评价指标。在假设的前提下统计了答案在不同窗口中的分布比例,具体结果见表 4。

表 4 答案在不同窗口中的分布

窗口	1	2	3	4	4 +
比例	51.4%	71.5%	84.0%	92.7%	7.3%

从表4 中可以发现窗口 4 中已包含了 92.7% 问题的答案,答案分布在窗口 4 以外的仅有 7.3%。有些问题可能回答得较晚,也可能无人回答。此外,部分问句的距离(其在文本中的顺序)较近,这使得各个问题的候选答案可能有交集。在窗口为 4 的前提下,本文选取了有答案的 2 643 个问题和对应候选答案作为实验数据,其中 2 000 组为训练集,643 组为测试集。

4.3　实验结果

问答对获取结果如表 5 所示。

表 5　问答对获取结果

窗口	方法	P	R	F1
4	Rank	0.656	0.656	0.656
	Re_rank	0.677	0.677	0.677
	Re_rank_com	**0.690**	**0.690**	**0.690**

从表 5 我们可以得出:

候选答案的次序对问答对获取有明显的影响,考虑次序加权的候选答案排序 Re_rank 比 rank 提升了 2.1 百分点。

使用共现词策略化解 Rank 与 Re_rank 的冲突是有效的,Re_rank_com 的各项指标均比 Re_rank 有了明显提升。

对问答对获取失败的数据进行统计分析,总结了两个主要因素:(1) 答案的初始得分过低。(2) 答案的回答顺序过于靠后。答案的初始得分过低时,即便其回答顺序靠前也无法通过 Re_rank 获得较大的"加分";答案的回答顺序靠后时,会在 Re_rank 中丧失原本的得分优势,等同于被"减分";若一个答案同时受此二因素影响,几乎都无法得到正确匹配。

5　结　语

经过本文的分析与研究可得到如下结论:(1) 基于规则的问句识别方法所用规则是有效的,具有较高的识别准确率。(2) 问句识别任务可以当作文本分类任务来处理,且 CNN 网络和 LSTM 网络可以较好地处理此类问题。(3) 规则与深度学习相结合的问句识别方法具有较强的鲁棒性,其召回率较单一方法均有较大提升,可以适应更丰富的任务类型。(4) 回答顺序在社交群问答对获取中有重要意义,合理使用此特征可提升问答对获取质量。

本文的工作仍有改进余地,未来可以适当考虑省略、指代等因素对问答对获取的影响。此外,可以考虑使用共现词之外的特征化解 Re_rank 与 Rank 的冲突,进一步提升问答对获取质量。

参　考　文　献

[1] 孙林. 基于在线论坛的问答对识别研究与问答系统实现[D]. 哈尔滨:哈尔滨工业大学,2010:12 - 15.

[2] 王宝勋. 面向网络社区问答对的语义挖掘研究[D]. 哈尔滨:哈尔滨工业大学,2013:40 - 47.

[3] 刘佳宾,胡国平,陈超,等. 基于决策树和马尔可夫链的问答对自动提取[J]. 中文信息学报,2007,21(2):46 - 51.

[4] 孟祥燕,余正涛,许洋波,等. 基于改进贝叶斯的领域问答对自动获取[J]. 广西师范大学学报(自然科学版),2009,27(1):189 - 192.

[5] 周飞云. 如何识别是设问句还是反问句? [J]. 中文自学指导,2000(2):41.

[6] 殷树林. 现代汉语反问句研究[D]. 福州:福建师范大学,2006:40 - 47.

[7] 文治,李旸,王素格,等. 融合反问特征的卷积神经网络的中文反问句识别[J]. 中文信息学报,2019,33(1):68 - 76.

[8] 熊作平. 汉、英否定是非问句及其答语对比研究[D]. 成都:四川师范大学,2008:33 - 36.

[9] 陈彩霞. 语气词、语调和标点在留学生是非问句识别中作用的实验研究[D]. 广州:暨南大学,2017:24 - 29.

[10] 侯永帅,张耀允,王晓龙,等. 中文问答系统中时间敏感问句的识别和检索[J]. 计算机研究与发展,2012,(12):2612 - 2620.

[11] 金丽娇,傅云斌,董启文. 基于卷积神经网络的自动问答[J]. 华东师范大学学报(自然科学版),2017(5):66 - 79.

[12] 荣光辉,黄震华. 基于深度学习的问答匹配方法[J]. 计算机应用,2017,37(10):2861 - 2865.

[13] 邢超. 智能问答系统的设计与实现[D]. 北京:北京交通大学,2015:20 - 26.

[14] Shen Y K, Rong W G, Sun Z W, et al. Question/answer matching for CQA system via combining lexical and sequential information[C]//Proceedings of 29 AAAI Conference on Artificial Intelligence,2015:275 - 281.

[15] Yang X, Khabsa M, Wang M S, et al. Adversarial training for community question answer selection based on multi-scale matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2019:395 - 402.

[16] Wu B, Wei B, Liu J, et al. Facet annotation by extending CNN with a matching strategy[J]. Neural Computing,2018,30(6):1647 - 1672.

[17] Lima E, Shi W S, Liu X M, et al. Integrating multi-level tag recommendation with external knowledge bases for automatic question answering[J]. ACM Transactions on Internet Technology. 2019,19(3):1 - 22.

[18] Li P, Li W, He Z Y, et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering[EB]. arXiv:1607.06275, 2016.