面向低资源场景的领域知识图谱问答系统

张敏[1]，刘宇嘉[2]，陈一萌[1]，陶仕敏[1]

[1] 华为技术有限公司2012实验室，北京，100095
[2] 中国人民解放军国防科技大学，湖南长沙，410073
{zhangmin186,chenyimeng}@huawei.com

摘要 本文介绍了我们在CCKS-2021“运营商知识图谱推理问答”任务上的技术方案。该方案由实体识别和链接、答案类型预测、属性名预测、约束抽取和消歧、答案查询五个模块级联组成。由于任务允许使用的训练数据非常有限，我们设计了面向低资源场景的三阶段实体识别和链接方法、基于微调预训练语言模型的答案类型预测和属性名预测方法、以及基于规则的约束抽取和消歧及答案查询方法。本文方法在复赛测试集上取得了0.9863的F1值，排名第一，其在低资源场景下的有效性得到验证。

关键词: 知识图谱问答，低资源场景，实体识别和链接，预训练语言模型

1 引言

知识图谱（Knowledge Graph，KG）[1]是以图的形式表示客观世界中的实体（概念、人、事物）及其之间关系的知识库。早在2006年，语义网[2]就定义了类似的概念，呼吁推广、完善使用本体模型来形式化表达数据中的隐含语义；2012年谷歌[3]正式提出知识图谱的概念，其初衷是为了提升搜索引擎的结果相关性和用户体验。目前，随着智能信息服务应用的不断发展，知识图谱已广泛应用于智能搜索、智能问答和个性化推荐等领域[4]。

知识图谱问答（Knowledge Based Question Answer, KBQA）[5]随着深度学习技术[6]在自然语言处理领域[7]的不断发展和突破，又因其结构化知识便于管理和维护，已成为智能问答领域研究和应用的热点。目前KBQA的主流方法可分为两类：1）基于语义解析（Semantic Parser）的方法，将自然语言转化成中间的语义表示（Logical Forms），然后再将其转化为可在KG中执行的描述性语言（如SPARQL语言），该类方法往往需要大量的人工标注数据，可在限定领域和小规模知识库上取得较好的结果；2）基于信息检索（Information Retrieval）的方法，首先会确定用户问题中的实体提及词(Entity Mention)，然后链接到KG中的主题实体（Topic Entity），并将与Topic Entity相关的子图（Subgraph）提取出来作为候选答案集合，然后分别从用户问题和候选答案中抽取特征，最后利用排序模型对用户问题和候选答案进行建模并预测，该类方法不需要人工定制规则且能够扩展到更大更复杂的知识库上。

我们参与了CCKS-2021“运营商知识图谱推理问答”任务，提出了基于语义解析的知识图谱问答系统，包括实体识别和链接、答案类型预测、属性名预

测、约束抽取和消歧、答案查询五个模块，并针对该任务的低资源场景特性设计了相应的技术方案，在复赛测试集上到达了0.9863的F1值，排名第一。接下来，将对任务、技术方案和实验进行介绍。

2 任务介绍

给定运营商知识图谱schema和三元组数据，以及5000条标注数据（用户问题、答案类型、属性名、实体、约束属性名、约束属性值、约束算子、答案），在仅可使用公开数据资源的条件下，推理计算出测试集中用户问题的答案。

3 技术方案

本文的技术方案由实体识别和链接、答案类型预测、属性名预测、约束抽取和消歧、答案查询五个模块级联组成，示例如图1所示。其中，实体识别和链接模块负责抽取用户问题（Query）中的实体并链接到知识库中的实体，答案类型预测模块负责判断Query的句子类型（属性值、并列句、比较句），属性名预测模块根据不同的Query句子类型进行知识库中的属性名预测，约束抽取和消歧模块根据识别到实体对应的知识库属性对Query生成正确的约束条件，答案查询模块根据上述模块输出结果（实体、答案类型、属性名、约束条件）查询知识库计算出最终结果。

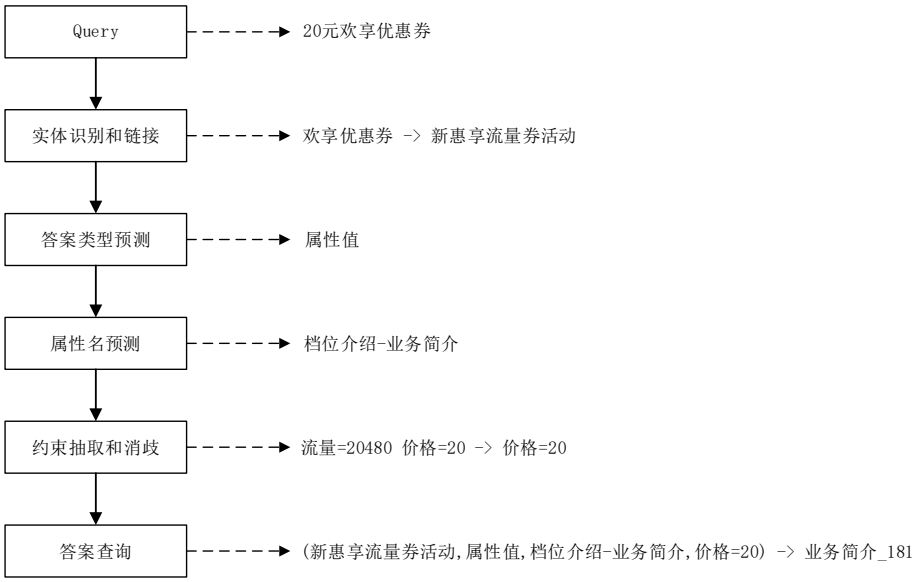

图1. 技术方案示例图

3.1 数据

根据任务要求，本文使用官方提供的5000条标注数据和公开预训练模型Chinese BERT[1]进行相关模型的训练和测试，并根据标注数据对官方提供的实体同义词字典进行了扩充。对用户Query数据，进行了常规文本预处理：大写转小写、繁体转简体、全半角转换、中文数字转阿拉伯数字、纠错。

3.2 实体识别和链接

由于训练数据有限，本文提出了一个面向低资源场景的三阶段实体识别和链接方法，如图2所示。

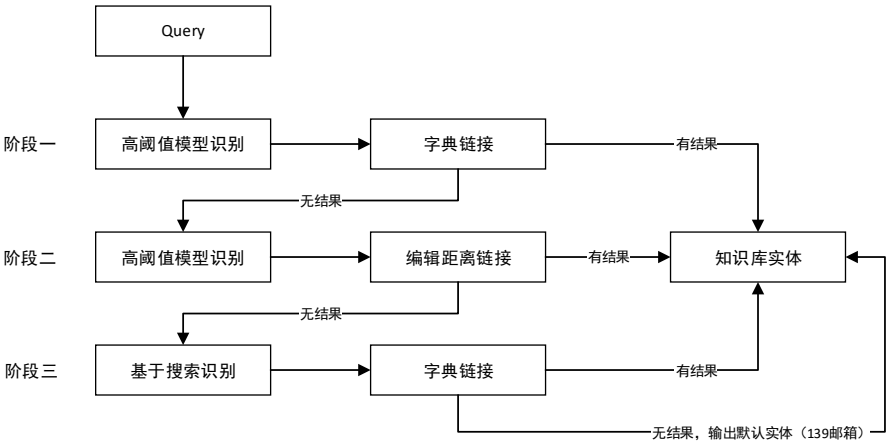

图2. 三阶段实体识别和链接方法

每个阶段的具体描述如下：

- 阶段一“高阈值模型识别 + 字典链接”：基于BERT+CRF在标注数据上训练实体识别模型，预测时仅考虑似然概率高于指定阈值（实验中取值0.925）的实体词，若该实体命中实体同义词字典，则输出对应的知识库实体，否则进入阶段2。
- 阶段二“高阈值模型识别 + 编辑距离链接”：计算模型实体词与同义词字典中所有实体词的编辑距离和相似度，若编辑距离和相似度满足指定阈值（实验中分别取2和0.4），则输出按编辑距离和相似度排序下的最优实体词对应的知识库实体，否则进入阶段3。其中，相似度的计算方式如下：

$$\text{相似度} = 1 - \frac{\textit{编辑距离}\left(\textit{模型实体词},\textit{字典实体词}\right)}{max\left(\textit{模型实体词长度},\textit{字典实体词长度}\right)} \quad (1)$$

[1] https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

- 阶段三“基于搜索识别 + 字典链接”：对实体同义词字典建立单字倒排索引，使用Query中每个单字查询索引获得的并集作为候选实体词集合，并按公式（2）计算候选实体词得分，输出最高得分实体词对应的知识库实体。

$$\text{Score} = \begin{cases} match\_length + match\_rate, if\ match\_rate > 0.5 \\ 0.5 * (match\_length + match\_rate), otherwise \end{cases} \quad (2)$$

匹配长度$match_length =$ *最大公共子序列长度*$(Query,$ *候选实体词*$)$，匹配度$match_rate = match_length/$ *候选实体词长度*。

需要说明的是，阶段三中，若匹配长度低于阈值或匹配内容过于宽泛，则输出默认实体词（实验中为“139邮箱”）。

3.3 答案类型预测

任务中Query的答案类型有3种（属性值、并列句、比较句），本文使用标注数据在基于BERT的多分类模型[7]上进行微调，从而实现对Query的答案类型预测。考虑到可用的标注数据量较少且实体词内容与答案类型预测无关，因此将Query中的实体词统一替换为_entity_，让模型更好地去学习与答案类型相关的字词，示例如下：

Query = 如何开通亲情号 变换为 *Query = 如何开通_entity_*

3.4 属性名预测

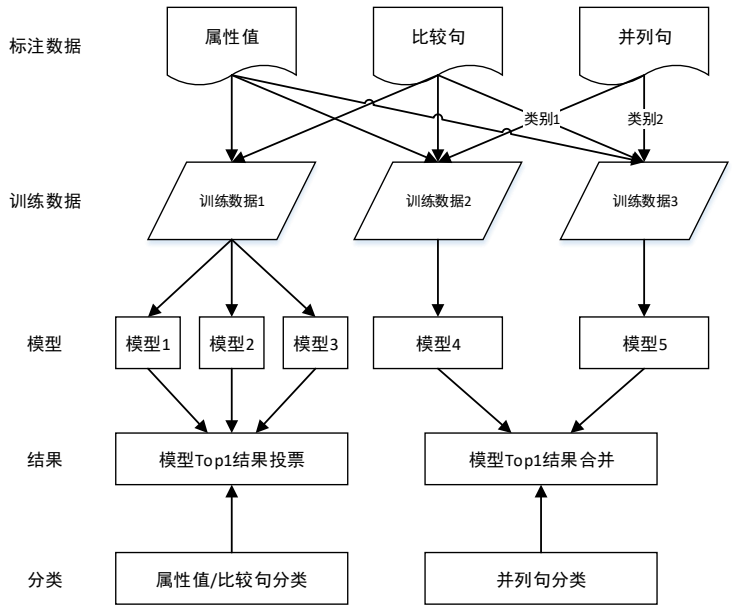

图3. 属性名预测流程图

官方提供的知识图谱schema文件中共出现了25类属性，而提供的人工标注数据只出现了19类，因此本文根据schema文件中的例句pattern，对剩余6类进行了数据扩充，以保证属性名覆盖的完整性。

从Query的句子类型来看，属性值和比较句仅需要预测1个属性名，而并列句需要预测2个属性名，因此本文为其分别训练模型，其中对Query中的实体词采用了与3.3节相同的替换方式，整体流程如图3所示。

属性值/比较句的属性名预测

由于仅需要预测1个属性名，我们仍采用微调BERT多分类模型，考虑到可用的训练数据较少（训练数据为“属性值”和“比较句”类型的标注数据），为防止模型过拟合，微调Epoch次数控制在20以内；另一方面，为保证在测试数据上获得更高的准确率，我们通过设置不同的随机数种子训练出多个模型进行集成（实验中训练了3个模型进行投票）。

并列句的属性名预测

并列句需要预测2个属性名，本质上是一个多标签多分类问题，本文通过训练2个单标签多分类模型实现2个属性名的预测。其中，模型的训练数据构造方式为：标注数据中的属性值和比较句均进入2个模型的训练集；将标注数据中的每条并列句拆分为2条属性值句子，分别进入2个模型的训练集。在预测并列句属性名时，合并2个模型的Top 1预测值作为结果，若2个模型的Top 1结果相同，则选择第1个模型的Top 2预测值作为结果。

3.5 约束抽取和消歧

约束条件的抽取和消歧是在完成Query实体识别和链接后进行的，具体流程如图4所示。

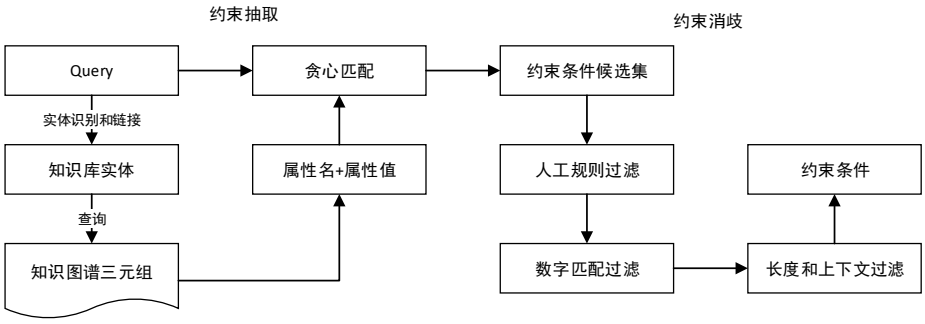

图4. 约束抽取和消歧流程图

约束抽取

取出Query对应知识库实体所有的“属性名”和“属性值”，按字符串方式匹配方式检查Query是否包含“属性值”，如果包含，则作为约束条件引入。另外，max和min形式的约束条件则直接根据Query是否包含相关的关键词引入。本文使用的约束抽取方式，目的是尽可能保证约束条件的召回。

约束消歧

由于约束抽取时以最大召回为目标，因此会存在误召回的问题，需要通过以下3个阶段的过滤和消歧来解决。

- 阶段一：基于人工规则的约束条件过滤
  从标注数据的约束条件中总结出的规则，如“产品名称”不能作为约束条件，实验中共总结出4条过滤规则。
- 阶段二：基于数字匹配的约束条件过滤
  由于约束抽取时仅考虑了Query是否包含“属性值”，因此会存在匹配上多个数字类型约束条件的情况，如Query为“*15元留言信箱怎么购买*”，“属性值”为1、5、15的约束条件都可以匹配上，因此针对数字匹配的情况，通过正则过滤数字部分匹配的约束条件。
- 阶段三：基于长度和上下文的约束条件消歧
  对约束值在Query中存在重叠的约束条件，如Query为“*怎么用半年包短信发送提醒*”会匹配上“半年包”和“年包”这2个约束条件，则保留长度更长的约束条件；在运营商场景下，“价格”和“流量”约束条件存在歧义的可能最大，因此本文针对性地根据此类约束值在Query中的上下文信息（相邻共现词、属性名）来进行消歧，如Query为“*我想开通700流量加油包需要花多少钱*”中的“700”究竟是流量约束还是价格约束，结合Query的属性名为“档位介绍-价格”，过滤价格约束。

由于标注数据量较少，本文的约束抽取和消歧都基于规则方法，具备较好的泛化能力，并在实验中得到验证。

3.6 答案查询

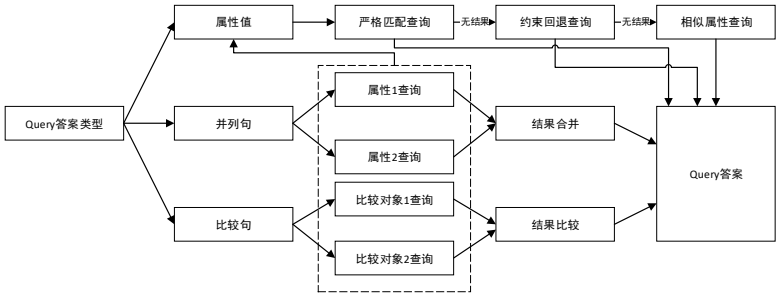

图5. 答案查询流程

本文对官方提供的运营商知识图谱三元组数据建立了倒排和正排索引，以实现在得到上述模块处理结果后查询出Query的答案。在查询答案时，本文根据答案类型进行了分别处理，如图5所示，具体描述如下：

类型1：属性值

1. 先按实体、约束条件、属性名均严格匹配方式在索引中查询结果，有结果则直接返回，否则进入步骤2
2. 约束回退：减少约束条件的个数进行查询（最多减少2个约束条件），有结果直接返回，否则进入步骤3
3. 相似属性名：使用相似属性名进行查询（如将“开通方式”改为“开通条件”），约束条件与步骤1相同

类型2：并列句

1. 将并列句拆分为2个属性值分别查询结果
2. 合并查询结果并返回

类型3：比较句

1. 将约束条件分为2个部分：比较对象和正常的约束条件
2. 按属性值方式查询每个比较对象的结果
3. 根据Query比较方式（是否、大于、小于）比较步骤2中的查询结果并返回

至此完成了本文面向低资源场景的领域知识图谱问答技术方案的介绍，其采用了模型+规则的模块级联方式，可保证较好的泛化能力。

4 实验

CCKS-2021“运营商知识图谱推理问答”任务组织方提供了2867条知识图谱三元组、138个知识库实体及其551个同义词，5000条标注数据以及1000条初赛未标注数据。在模型训练过程中，实体识别模型将5000条标注数据按9:1划分为训练集和测试集，答案类型预测和属性名预测模型使用了全部标注数据进行训练并以1000条初赛未标注数据作为测试集，所有模型参数均采用默认设置。另外，通过对知识图谱Schema文件和标注数据的分析，我们增加了6个属性名类别共22条训练数据和25个知识库实体同义词。需要说明的是，对标注的5000条数据，我们过滤了标注实体及其同义词均不在Query中出现的数据，共249条，即用于模型训练的标注数据为4751条。

4.1 模型效果评估

本节给出实体识别模型、答案类型预测模型、属性名预测模型的实验结果，如表1和表2所示。

表1. 实体识别模型和答案类型预测模型实验结果

模型名称	训练集大小	Epoch次数	测试集F1值
实体识别模型	4304	32	0.976
答案类型预测模型	4751	4	0.995

表2. 属性名预测模型实验结果

答案类型	模型名称	训练集大小	Epoch次数	测试集F1值
属性名/比较句	模型1	4082	18	0.990
	模型2	4082	16	0.996
	模型3	4082	8	0.992
	集成	--	--	0.999
并列句	模型1	4773	10	--
	模型2	4773	10	--
	合并	--	--	0.998

其中，答案类型预测模型和属性名预测模型的测试集都来自初赛1000条未标注数据（其标注来自线下训练多个模型的投票结果，对不一致的投票结果采用人工校验或初赛线上校验方式进行修正），因此其F1值会存在高估。

4.2 端到端效果评估

本节将给出本文知识图谱问答系统端到端的效果评估，在初赛测试集和复赛测试集上的评估结果如表3所示。其中，初赛测试集的标注与4.1节方式相同。

表3. 端到端效果评估

测试集名称	测试集大小	F1值
初赛	1000	0.9880
复赛	1000	0.9863

5 总结

本文针对“运营商知识图谱推理问答”任务提出了一种面向低资源场景的领域知识图谱问答系统，其由实体识别和链接、答案类型预测、属性名预测、约束

抽取和消歧、答案查询五个模块级联组成，在小数据量场景下以模型+规则方式取得了较好的泛化能力，系统在复赛测试集上到达了0.9863的F1值。

参考文献

1. 肖仰华, *知识图谱概念与技术*. 2020: 电子工业出版社.
2. Berners-Lee, T., J. Hendler, and O. Lassila, *The Semantic Web*. Scientific American Magazine, 2008. **23**(1): p. 1-4.
3. Singhal, A., *Introducing the Knowledge Graph: things, not strings*. 2012, Official Blog of Google.
4. 徐增林, et al., *知识图谱技术综述*. 电子科技大学学报, 2016. **45**(4): p. 589-606.
5. Cui, W., et al., *KBQA: Learning Question Answering over QA Corpora and Knowledge Bases*. Proceedings of the VLDB Endowment, 2017. **10**(5): p. 565-576.
6. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the Dimensionality of Data with Neural Networks*. Science, 2006. **313**(5786): p. 504-507.
7. Devlin, J., et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in *NAACL*. 2019, Association for Computational Linguistics. p. 4171–4186.