湖南农业大学学报(自然科学版) 2025，51(3)：97–109. DOI：10.13331/j.cnki.jhau.2025.03.013
Journal of Hunan Agricultural University(Natural Sciences)

引用格式：
陈婷，朱昌群. 融合知识图谱和语义信息的烟叶分级问答系统[J]. 湖南农业大学学报(自然科学版)，2025，51(3)：97–109.
CHEN T，ZHU C Q. Tobacco grading question answering system integrating knowledge graph and semantic information[J]. Journal of Hunan Agricultural University(Natural Sciences)，2025，51(3)：97–109.
投稿网址：http://xb.hunau.edu.cn

融合知识图谱和语义信息的烟叶分级问答系统

陈婷，朱昌群

(昆明理工大学机电工程学院，云南 昆明 650500)

摘要：针对烟叶分级领域知识冗余且没有专业化平台用于学术检索的现状，采集多源烟叶分级数据并结合自顶向下的方法构建烟叶分级知识图谱，并以此为基础开发智能问答系统。其核心技术主要包括：1) 采集烟叶分级数据，经过命名实体识别(NER)以及关系抽取(RE)后提取三元组信息，并将其导入Neo4j平台储存；2) 对于问句语义解析，采用融合图谱数据的BERT-BiGRU-MHSA-CRF模型提升问句实体识别效果，同时将自注意力机制融入BERT-TextCNN模型中，用于解析用户分级意图，再通过匹配模板并替换槽位信息以便自动化构建cypher查询语句，在Neo4j知识库中查询最精确的答案并返回。结果表明：构建的知识图谱包含6 620个实体，超过14 000条关系；基于问句实体识别模型BERT-BiGRU-MHSA-CRF的调和平均值F_1为94.12%，分级意图识别模型BERT-TextCNN-Attention的F_1为98.77%。综上，该系统实现了对烟叶分级相关的多类问题的快速检索和精确回答，可以为分级人员提供辅助。

关键词：领域知识图谱；语义解析；问答系统；烟叶分级；问句实体识别；意图识别

中图分类号：TS441 **文献标志码**：A **文章编号**：1007–1032(2025)03–0097–13

Tobacco grading question answering system integrating knowledge graph and semantic information

CHEN Ting, ZHU Changqun

(School of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: In view of the redundancy of knowledge in the field of tobacco grading and the absence of a professional platform for academic retrieving, the knowledge graph of tobacco grading was constructed by collecting multi-source tobacco grading data and combining the top-down method, and an intelligent question and answer system was developed on this basis. The core technologies are as follows. 1) Collecting tobacco leaf grading data through named entity recognition(NER) and relation extraction(RE) to extract triplet information, and import it into the Neo4j platform for storage. 2) For question semantic parsing, the BERT-BiGRU-MHSA-CRF model fused with graph data was used to improve the entity recognition effect of question sentences, and the self-attention mechanism was integrated into the BERT-TextCNN model to parse user hierarchical intent. Then, the cypher query statement was automatically constructed by matching the template and replacing the slot information, and the most accurate answer was retrieved and returned in the Neo4j knowledge base. The results showed that the constructed knowledge graph contains 6 620 entities and more than 14 000 relationships. The harmonic mean F_1 of the question entity recognition model BERT-BiGRU-MHSA-CRF was 94.12%, and the F_1 of the hierarchical intent recognition model BERT-TextCNN-Attention was 98.77%. In summary, the system can quickly retrieve and accurately answer multiple types of questions related to tobacco grading, which can

收稿日期：2024–01–23 **修回日期**：2025–03–05
基金项目：国家自然科学基金项目(61761024)
作者简介：陈婷(1971—)，女，云南昆明人，硕士，副教授，主要从事企业集成及信息化研究，kmct@163.com

provide auxiliary functions for graders.

Keywords: domain knowledge graph; semantic analysis; question answering system; tobacco grading; question entity recognition; intent recognition

烟叶分级[1]是烟草工业中的关键环节，不仅影响烟农的切身利益，还会对烟草税收造成一定的影响。随着工业信息化的发展，烟叶分级领域也受到了影响，出现了很多自动化分级的机器，不用依靠人工就可以完成等级鉴定。但大多数机器仅能实现单片叶分级，导致分级效率低，需要大量人工和分级设备参与分级工作，不利于烟叶分级领域的发展。目前，烟叶分级工作仍需烟叶分级员的参与，这就要求其具备充分的烟叶分级知识，但是烟叶分级是一项复杂、影响因素较多的工作，并且没有专业化的平台可供学习了解，存在数据冗余现象。随着互联网的发展，人们对各种知识的了解越来越依赖网络，而智能问答[2]也逐渐成为一种新型的人机交互工具。烟叶分级领域问答是一种限定域的问答方式，其核心在于数据存储与语义解析。由于烟叶分级领域的知识过于专业化，且多为半结构化和非结构化数据，使用传统的关系数据库无法充分挖掘、利用烟叶分级知识，因此，本文采用知识图谱[3]的方式对数据进行存储。知识图谱本质上是一种语义网络，它采用图结构的方式对数据进行存储，优质的图谱数据可以提升领域智能问答系统的整体效果。语义解析[4]主要包括问句实体识别和问句意图识别，其核心思想是将用户提出的自然语句中的重要实体信息和意图信息提取出来，以实现答案检索，语义解析的效果会直接影响整个问答系统的性能。为提升图谱质量和语义解析效果，本文针对烟叶分级专业化书籍、文献和培训资料数据，经过命名实体识别任务和关系抽取任务，自动抽取数据中的三元组信息并将其存储在Neo4j平台，构建可视化的烟叶分级领域知识图谱，并在此基础上提出融合前期图谱构建数据的问句实体识别模型和解决分级多类问题的意图识别模型，以提升语义解析效果，从而实现烟叶分级领域智能问答系统的构建。

1 基于烟叶分级领域知识图谱的问答系统

基于烟叶分级领域知识图谱的问答系统主要包括图谱构建和语义解析2个模块。知识图谱是以图结构的形式来存储数据，其中节点表示实体，节点之间的边表示关系，作为一种图形化的存储工具，它更适用于知识问答、语义检索等任务。目前，领域知识图谱[5]的研究集中于医疗、农业、动植物等领域，其中网络资源是其主要数据来源。刘燕等[6]利用医学百科知识和电子病历文本构建了医疗知识图谱；QIN等[7]通过自动与人工双模式并结合联合抽取信息的方法构建了有关农业信息方面的知识图谱；田梦晖等[8]通过Albert模型进行知识抽取，成功构建了珍稀濒危植物领域的知识图谱。近年来，知识图谱所涉及的领域越来越多，但针对烟叶分级领域知识图谱的研究仍然较少，主要原因是烟叶分级知识复杂，评价烟叶分级的指标较多，部分等级之间的指标区分度较小，难以实现精准化的分级；此外，烟叶分级领域没有专业化的网站用于爬取分级知识，需要从专业书籍、资料中抽取信息。

语义解析部分主要包括命名实体识别和意图识别(IR)两个部分。命名实体识别[9]主要是提取用户自然问句中的重要实体信息，如烟叶类型、产地、分级要素等，而意图识别[10]主要是关注用户的整体目的和意图，如查询产地信息、烟叶类型等。本文在命名实体识别和意图识别中融入槽位填充[11]等模块，可以更深层次地提取问句语义信息并进行处理，以达到构建符合用户意图的cypher查询语句，从而提升问答整体效果的目的。

命名实体识别(NER)的研究方法有基于词典和规则的方法[12]、基于传统机器学习的方法[13]以及深度学习方法[14]这3种。基于词典和规则的方法需要该领域的专业人士制定规则并设置大量的词典信息，这种方法需要大量的人工参与，泛化性能差；基于传统机器学习的方法需要大量的样本数据作为基础，然后通过概率统计模型对数据进行分析和特征提取。目前使用的最多的是深度学习方法，AN等[15]提出一种多头自注意力机制融入BiLSTM-CRF模型的方法，大幅度提升了模型在临床实体识别方面的精度；陈娜等[16]引入基于Transformer的双向编码器表征(BERT)，结合双向门控循环单元–条

件随机场(BiGRU-CRF)模型用于中文电子病历方面的实体检测并取得了较好的效果；YANG等[17]提出了BERT-MBiGRU-CRE模型以提高命名实体识别的精度，该模型采用多层双向门控循环单元取代双向长短期记忆网络，有效地提取到了全局上下文的语义特征。

意图识别(IR)本质上是一种句子级的分类任务，其目的是解析自然问句中的语义信息，更好地构建关系和属性查询，理解用户意图。目前使用得最多的意图识别方法也是深度学习的方法。WU等[18]提出了一种融合顺序信息和句子结构特征的意图识别模型，首先通过卷积神经网络(CNN)捕获文本中的局部显著特征，然后运用BiLSTM在局部上下文中提取顺序信息，最后，通过BERT处理以提取句子结构特征，该方法有效地提升了意图识别的准确率。余建明等[19]使用预训练模型ALBERT结合RE2模型并融入字词向量和编码向量对电网意图进行了有效识别；郭旭超等[20]则将意图识别联合槽位填充模块应用于农业病虫害的知识问答中，取得了较好的结果。

目前基于领域知识图谱的问答系统还存在以下问题：1) 数据大多来源于网络爬取，难以保证构建图谱的质量；2) 知识图谱仅作为知识问答的语料库，其与语义解析部分的关联度较低，导致图谱知识未被充分挖掘；3) 语义解析结果会直接影响问答系统整体效果，但语义解析可能会有偏差，如实体识别错误导致难以正确构建查询语句，降低问答系统性能。针对上述问题，本文采集烟叶分级领域专业化数据，并通过完整的流水线式图谱构建方法[21]以保证图谱质量，并在该图谱的基础上采用融合前期图谱构建数据的BERT-BiGRU-MHSA-CRF问句实体识别方法，充分挖掘图谱知识，同时结合BERT-TextCNN-Attention分级意图识别模型提升问句解析效果，以实现用户对烟叶分级知识的有效问答。

2　系统架构及技术实现

2.1　问答系统整体架构

本文所设计的问答系统总体框架如图1所示，主要包括知识图谱构建模块、知识问答模块和业务展示模块。首先，采集非结构化数据(专业书籍、培训资料)以及半结构化数据(烟叶分级文献)，对其进行数据清洗；然后，设定实体关系类型并进行标注；通过知识图谱构建模块将处理好的数据经命名实体识别任务和关系抽取任务处理以保证数据质量，最后抽取三元组信息并将其储存至Neo4j平台。

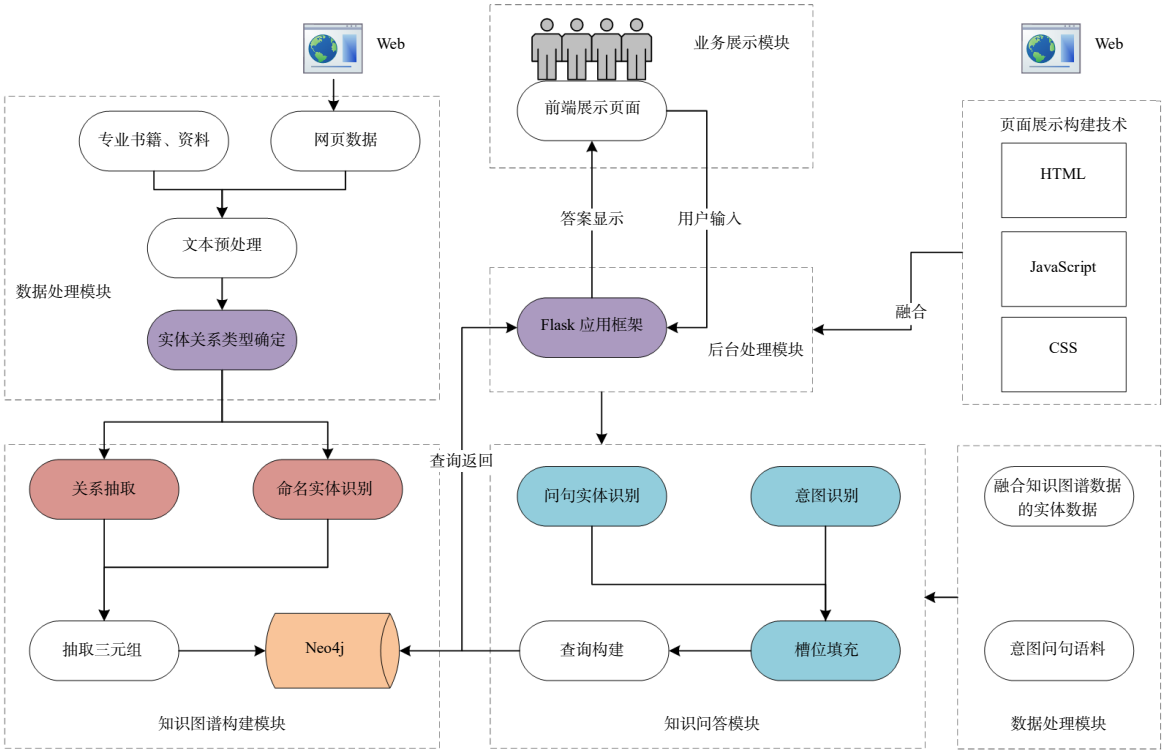

图1　问答系统总体框架

Fig.1　Overall framework of question answering system

知识问答模块的主要功能是理解用户的查询意图，即由用户输入自然问句,通过问句实体识别模型提取问句中的重要实体信息,再经过意图识别模型进一步解析用户的真正意图,最后将问句实体识别和意图识别的结果匹配至相应的模板进行槽位填充，再将槽位填充的结果自动转化为cypher查询语句，映射至Neo4j平台后查询答案并返回。后台处理模块采用Flask框架结合JavaScript以及HTML等技术以实现问答系统的可视化。

2.2 知识图谱构建

烟叶分级知识图谱的构建是烟叶分级问答系统的基础，它为问答系统提供了数据支撑，常规的问答系统大多采用爬虫技术爬取网络资源，经过数据清洗及相关预处理后再直接抽取三元组进行存储以供问答系统使用，目前尚无专业化的烟叶分级知识网络资源，因此，本文先通过流水线式的图谱构建方法将专业书籍、资料以及烟叶分级文献知识整理标注，然后通过命名实体识别任务和关系抽取任务处理数据，最后抽取三元组并储存至Neo4j平台，实现烟叶分级知识图谱的可视化。

2.2.1 命名实体识别

命名实体识别是自然语言处理(natural language processing，NLP)中的一个重要任务，其目标是从文本中识别和分类命名实体。本文根据烟叶分级知识的特殊性以及烤烟国家标准[22]，从烟叶类型、分级指标、识别方法以及其他描述烟叶分级的知识将烟叶分级数据划分为20类实体类型，具体信息如表1所示。

表1 烟叶分级实体类型

Table 1 Physical type of tobacco classifying

实体类型	中文含义	举例	实体类型	中文含义	举例
type	烟叶类型	烤烟、香料烟等	number	数字	8个正组等
pos	部位	顶叶、上二棚等	stats	属性	专有名词的描述
color	颜色	橘黄、柠檬黄等	location	地点	云南、四川等
group	组别	中部橘黄组等	organization	组织	烟草质监站等
tge	分级要素	成熟度、油分等	time	时间	1992年等
tgei	分级要素指标	香气量、吃味等	file	文件	《烤烟分级标准》等
tgeiv	分级要素指标值	有、稍有、少等	name	姓名	具体的人名
grade	等级	中部橘黄1级等	variable	变量	含水率等
way	识别方法	手摸、眼观等	des	代号	C1L、B等
effect	识别效果	强烈的刺激性等	method	分级方法	贝叶斯判别等

本文根据表1的实体划分，采用doccano标注工具进行数据标注并按BIO的数据格式输出，用于模型训练。本文所提出的BERT-BiGRU-MHSA-CRF模型使用了特征提取层与多头自注意力机制，使该模型相较于LSTM模型更加轻量化，模型的计算速度和识别效果有所提升，且较好地解决了长序列问题和模型输入特征关注度不够的问题。

2.2.2 关系抽取

关系抽取(relation extraction，RE)任务即从文本中提取实体之间关系的任务，通常涉及识别文本中的实体并确定它们之间的语义关系。该任务旨在从非结构化文本中提取结构化的信息。本文根据上述设定的实体类型结合烟叶分级知识的特点，划分了16类关系类型，具体信息如表2所示。

表2 烟叶分级关系类型

Table 2 Relationship types of tobacco grading

关系类型	中文含义	举例	关系类型	中文含义	举例
sub	从属关系	<顶叶，从属于，上部叶>	print	发布关系	<烟草专卖局，发布，烤烟分级标准>
des	代号关系	<上部叶，代号，B>	print_time	发布时间	<烤烟分级标准，发布时间，1992年>
attribute_value	属性值	<烤烟，属性值，烤烟定义>	color	颜色	<青黄烟，颜色，青黄色>
index_value	指标值	<长度，指标值，40 cm>	effect	影响关系	<成熟度，影响，油分 >
quantity	数量关系	<正组，数量，8个>	rec_method	识别方法	<中部橘黄1级，识别方法，感官判断>
variable_value	变量值	<伤残百分比，变量值，25%>	rec_effect	识别效果	<手摸，识别效果，有油润感>
adopt	使用关系	<姓名，使用，神经网络>	stipulate	规定关系	<烤烟分级标准，规定，成熟度要求>
place	产地关系	<烤烟，产地，云南曲靖>	feature_des	特征描述	<上部橘黄1级，特征描述，成熟度的特征>

本文根据上述的关系类别划分，采用doccano标注工具进行数据标注并按<rel, object, subject, text>的数据格式输出，用于模型训练。本文提出的BERT-BiLSTM-Attention关系抽取模型主要以BERT预训练模型为基础提取文本特征，然后整合双向长短期记忆网络捕捉序列信息，同时结合位置嵌入技术实现对上下文信息和实体位置的有效建模，以更好地理解实体在文本中的位置关系。

关系抽取模型结构如图2所示，文本输入之后会经过特征提取层、BiLSTM层和输出层。首先，利用BERT对输入文本进行编码；其次，结合实体位置信息，通过双向长短期记忆网络对BERT输出进行建模；然后，使用自注意力机制对BiLSTM输出进行加权，以强调关键信息；最后，通过线性层将加权后的特征映射到关系分类空间并输出结果。

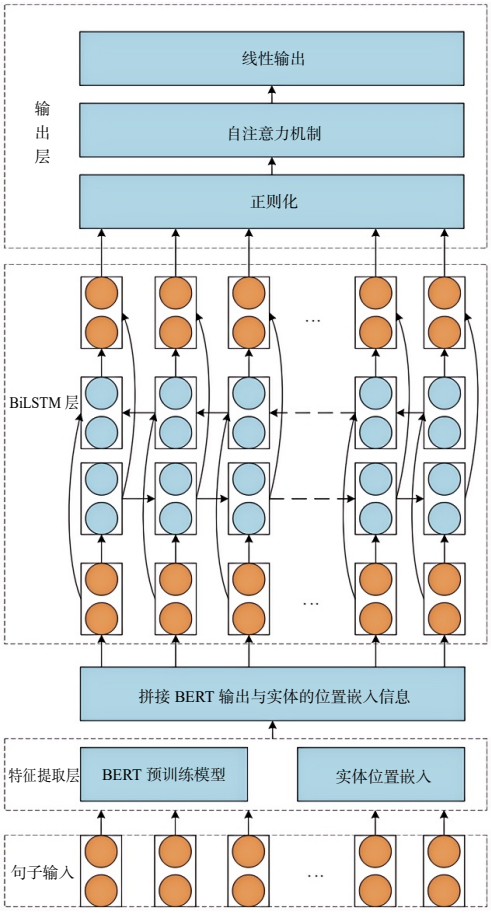

图2　关系抽取模型结构

Fig.2　Structure of relational extraction model

1) 特征提取层。为提升关系抽取准确率，在特征提取过程中添加实体位置信息。首先，将句子输入在嵌入层表示为初始化向量$\boldsymbol{I}=\{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n\}$，其中$\boldsymbol{t}_i$表示输入中第$i$个字符对应的特征向量，$\boldsymbol{I}$经过BERT层得到特征向量$\boldsymbol{H}_{\text{BERT}}=\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$，再将实体的起始位置和结束位置映射为向量$\boldsymbol{E}_{\text{entity}}=\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}$，其中，$\boldsymbol{E}_{\text{entity}}$表示实体位置的嵌入信息，$\boldsymbol{e}_i$是对应位置的向量；将BERT层的输出和实体位置信息的向量化结果按照时间步进行逐个拼接，得到结果$\boldsymbol{H}_{\text{concat}}=\{[\boldsymbol{h}_1;\boldsymbol{e}_1],[\boldsymbol{h}_2;\boldsymbol{e}_2],\cdots,[\boldsymbol{h}_n;\boldsymbol{e}_n]\}$，其中，$\boldsymbol{H}_{\text{concat}}$是拼接后的输入序列向量。

2) BiLSTM层。针对烟叶分级复杂的文本数据，采用BiLSTM捕捉输入序列的长距离依赖关系并提取上下文信息。将步骤1)中得到的拼接结果作为输入，经BiLSTM网络对序列信息进行建模后，得到输出$\boldsymbol{H}_{\text{BiLSTM}}=\{\boldsymbol{h}_1', \boldsymbol{h}_2', \cdots, \boldsymbol{h}_n',\}$，其中$\boldsymbol{H}_{\text{BiLSTM}}$是BiLSTM层的输出序列向量。

3) 输出层。引入正则化准则以防止过拟合，同时添加注意力机制用于加权组合BiLSTM的输出以及捕捉序列中的重要信息。

首先，对BiLSTM输出分别进行线性变换：

$$\boldsymbol{Q}=\boldsymbol{H}_{\text{BiLSTM}}\boldsymbol{W}_Q \tag{1}$$

$$\boldsymbol{K}=\boldsymbol{H}_{\text{BiLSTM}}\boldsymbol{W}_K \tag{2}$$

$$\boldsymbol{V}=\boldsymbol{H}_{\text{BiLSTM}}\boldsymbol{W}_V \tag{3}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为投影后的矩阵；$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的权重矩阵，

其次，计算注意力分数：

$$e_{ij}=\frac{\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d_k}} \tag{4}$$

其中：e_{ij}表示位置i对位置j的原始注意力分数，d_k为键/查询的维度。

然后，通过softmax函数对注意力权重进行归一化处理：

$$\partial_{ij}=\text{softmax}(e_{ij})=\frac{\exp(e_{ij})}{\sum_{k=1}^{n}e_{ik}} \tag{5}$$

其中，∂_{ij}表示位置i对位置j的归一化注意力权重；$\exp(e_{ij})$为指数函数，用于将分数转化为非负值。

随后，由∂_{ij}生成上下文向量：

$$\boldsymbol{C}_i=\sum_{j=1}^{n}(\partial_{ij}\boldsymbol{V}_j) \tag{6}$$

其中，$\boldsymbol{V}_j$表示输入序列中第j个位置通过权重矩阵

W_l变换后的特征表示。

最后，通过正则化准则以及softmax函数得到输出结果O：

$$O=\text{softmax}(\text{drop}(C)W'+b) \tag{7}$$

其中，C为自注意力机制输出的上下文表示矩阵，drop(C)表示对C的随机置零，W'为可学习的权重矩阵，b为线性层的偏置向量。

2.2.3 图谱可视化

在经过上述命名实体识别任务和关系抽取任务之后，充分验证了数据的有效性，保证了后续构建图谱的质量。通过python脚本从json文件格式的三元组数据中生成形如<object, object_type, relation, subject, subject_type>格式的数据，并将其保存为xlsx文件，然后调用py2neo库将实体和关系信息映射至Neo4j平台，生成的烟叶分级知识图谱如图3所示。

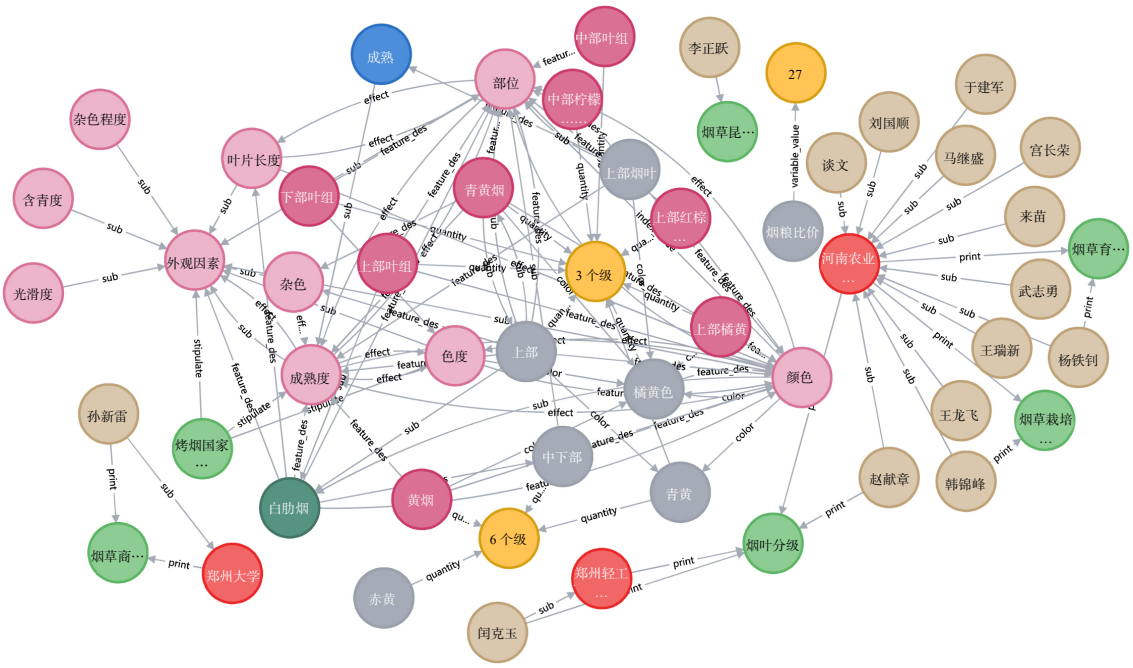

图3 烟叶分级知识图谱

Fig.3 Knowledge map of tobacco grading

2.3 问答系统构建

问答系统的构建可以合理地应用知识图谱中丰富的结构化语义信息，使得人机之间的交互更加高效。用户可以根据自己的意图对系统发起提问，系统解析用户问句，经过问句实体识别和多阶段意图识别后，根据解析结果匹配相应的模板并进行槽位填充生成cypher查询语句，在语义库中进行查询，最终将最佳答案反馈给用户。

2.3.1 问句实体识别

问句实体识别是自然语言处理的基础任务之一，也是实现问答系统的关键技术。在烟叶分级问答系统中，大多数问题的答案都与中心实体存在关联，因此，确定了问句中心实体就可以通过查询其子图找到问题答案。

本文设计融合多头自注意力机制的BERT-BiGRU-CRF模型以提升问句实体识别效果。为了充分利用图谱中的知识，将构建知识图谱的数据和生成的问答数据作为训练语料放入模型训练。问句实体识别模型结构如图4所示。训练语料经过BERT预训练层转化为特征向量，再由双向门控循环网络和Linear层对BERT层的向量输出进行文本编码；然后由线性层的权重矩阵整合BiGRU输出，再利用多头自注意力机制提取位置特征，使模型更加关注上下文语义信息；最后，通过CRF层引入转移矩阵对标签之间的依赖关系进行建模，以保证标签序列语法的合理性，同时输出问句预测标签。

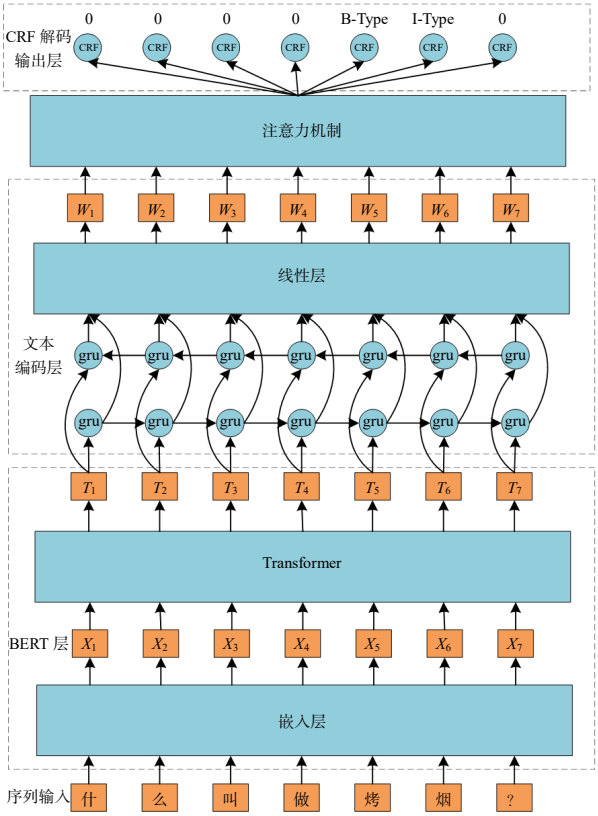

图4　问句实体识别模型结构

Fig.4　Structure of entity recognition model

以问句“什么叫做烤烟？”为例，该问句实体识别过程如下。

1) BERT层嵌入。BERT核心网络结构是由多层双向的Transformer Encoder构成，自然问句输入后由多个嵌入层进行编码相加得到[X_1, X_2,…, X_n]序列，再以Transformer Encoder结构作为特征提取器，增强语义特征的提取能力，生成特征词向量$\boldsymbol{T}$=[T_1, T_2,…,T_n]，即为BERT模型输出的词向量列表，其中$\boldsymbol{T}\in\mathbf{R}^{nd}$，$\mathbf{R}$表示实数集，$n$为文本序列长度，$d$为向量维度。

2) 文本编码层。文本编码层主要由双向门控循环单元(BiGRU)对BERT层输出的特征词向量进行全局特征提取。GRU是循环神经网络(RNN)的变体，与长短期记忆网络(LSTM)一样通过“门”结构控制信息传输，但相较于LSTM结构更简单，模型计算速率更快且模型效果与LSTM一致或更优。步骤1)中得到的特征向量$\boldsymbol{T}$传入文本编码层后由BiGRU利用双向的神经单元提取上下文特征，并对输出进行加权求和，再由线性层将d维向量映射为 m 维向量，得到文本编码层的输出标签向量列表$\boldsymbol{W}$=[W_1, W_2,…, W_n]，$\boldsymbol{W}\in\mathbf{R}^{nm}$，其中 m 为实体类型标签数。

$$\overrightarrow{\boldsymbol{W}_t}=G(\boldsymbol{x}_t,\overrightarrow{\boldsymbol{W}_{t-1}}) \tag{8}$$

$$\overleftarrow{\boldsymbol{W}_t}=G(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{W}_{t-1}}) \tag{9}$$

$$\boldsymbol{W}_t=[\overrightarrow{\boldsymbol{W}_t};\overleftarrow{\boldsymbol{W}_t}] \tag{10}$$

其中，$\boldsymbol{W}_t$为BiGRU当前时刻隐含层的输出状态，它由当前时刻t的输入$\boldsymbol{x}_t$、(t–1)时刻正向隐含层输出$\overrightarrow{\boldsymbol{W}_{t-1}}$和反向隐含层输出$\overleftarrow{\boldsymbol{W}_{t-1}}$共同决定，而[$\overrightarrow{\boldsymbol{W}_t}$;$\overleftarrow{\boldsymbol{W}_t}$]为当前时刻正向和反向隐含层输出矩阵水平拼接的结果。

3) 多头自注意力机制层。烟叶分级数据中有较多的长文本，如对烟叶等级的特征描述、识别方法的识别效果描述等，这些长文本会影响最终的识别效果。因此，本文引入多头自注意力机制层，加强对长文本的关注，减少对短文本的关注，更有效地获取上下文的语义特征，提升模型提取局部特征的能力，其具体实现过程如下。

由隐含层输出$\boldsymbol{W}_t$经全连接层得到$\boldsymbol{N}_t$，$\boldsymbol{N}_t$为当前信息与上下文信息相关的注意力权重向量。

$$\boldsymbol{N}_t=\tanh(\boldsymbol{W}_t\boldsymbol{H}_t+\boldsymbol{B}_t) \tag{11}$$

其中：$\boldsymbol{H}_t$为权重向量，$\boldsymbol{B}_t$为注意力层的偏置向量，tanh 为激活函数。

权重向量$\boldsymbol{N}_t$通过softmax函数的归一化处理后得到注意力分数向量$\boldsymbol{U}_t$：

$$\boldsymbol{U}_t=\text{soft max}(\boldsymbol{N}_t)=\frac{\exp(\boldsymbol{N}_t)}{\sum_{t=1}^{n}\exp(\boldsymbol{N}_t)} \tag{12}$$

双向门控循环单元输出$\boldsymbol{W}_t$经注意力机制权重分配后输出加权全局语义特征向量$\boldsymbol{S}_t$：

$$\boldsymbol{S}_t=\sum_{t=1}^{n}(\boldsymbol{U}_t\boldsymbol{W}_t) \tag{13}$$

4) CRF输出层。文本编码层和注意力机制的引入可以有效处理文本长距离依赖的问题，但无法处理标签之间的依赖关系。因此，本文引入条件随机场来捕捉标签之间的依赖关系，并对整个标签序列进行优化，最终由CRF层解码输出标签序列结果。其具体过程如下：由BiGRU和注意力机制输出得到得分矩阵$\boldsymbol{S}$，$\boldsymbol{S}\in\mathbf{R}^{nm}$，$S_{ij}$表示文本序列中第$i$个字符$x_i$的第$j$个标签分数。

首先，计算文本序列 $X=\{x_1,x_2,\cdots,x_n\}$ 的预测标签序列 $Y=\{y_1,y_2,\cdots,y_n\}$ 的得分。

$$\text{score}(X,Y)=\sum_{i=0}^{n}U_{y_i,y_{i+1}}+\sum_{i=1}^{n}S_{i,y_i} \tag{14}$$

其中，$\boldsymbol{U}$为转移分数矩阵，$\boldsymbol{U}\in\mathbf{R}^{(m+2)(m+2)}$，$U_{y_i,y_{i+1}}$为标签$y_i$转移到$y_{i+1}$的分数，$S_{i,y_i}$为输入文本序列第$i$个字符预测为标签$y_i$的概率。

然后，利用softmax函数进行归一化处理，计算输出标签序列Y的概率：

$$P(Y\mid X)\frac{\mathrm{e}^{\mathrm{score}(X,Y)}}{\sum_{Y'\in Y_X}\mathrm{e}^{\mathrm{score}(X,Y')}} \tag{15}$$

其中，Y_X为所有可能标签序列集合，Y'为真实标签序列。

最后，采用维特比算法得到文本序列X的全局最优标签序列Y^*，Y^*即为输出概率最大的标签集合。

$$Y^*=\arg\max_{Y\in Y_x}[\mathrm{score}(X,Y')] \tag{16}$$

例如，示例问句“什么叫做烤烟？”最终对应的输出标签序列为“['O', 'O', 'O', 'O', 'B-Type', 'I-Type','O']”，并从中抽取出实体信息“[['Type','烤烟']]”。

经过上述的流程，基于BERT-BiGRU-MHSA-CRF并结合图谱构建数据的问句实体识别模型可以准确地将自然问句中的实体信息抽取出来，以便于后续的槽位填充以及图谱查询构建工作。

2.3.2 意图识别

意图识别(IR)本质上是一种句子级的分类任务，其识别的准确性决定了图谱搜索的准确性和问答系统的智能性。本文设计的意图识别模型是将意图识别分为多阶段进行处理，其流程如图5所示。

由图5可见：问句实体识别任务判断初始输入的自然问句中是否有分级实体存在，若没有分级实体，则由初始意图模块使用BERT预训练模型进行语义相似度解析，返回概率值最大的标签，并根据设定的标签回复模板进行回复；若存在分级实体，则将实体信息融入分级意图识别模型，并将问句实体识别出的实体数量作为问句意图划分的一项特征。若实体数量为1，则将问句划入简单分级意图，再直接使用本文提出的BERT-TextCNN-Attention模型返回识别结果；若存在多个实体，则将其纳入复杂分级意图，再通过本文模型识别各项子类任务，分别处理各项子类任务后返回识别结果。例如，“根据特征判断大致烟叶等级”的相关自然问句中一般包含多个分级实体，因为烟叶分级是一个复杂的处理过程，它需要经过识别部位、颜色再结合其他特征才可以大致完成烟叶等级的判断，因此，将上述类型的复杂问题分为几个子任务即先识别问句中的部位特征并判断部位，再识别颜色特征并判断颜色，最后结合其他特征初步判别烟叶等级。

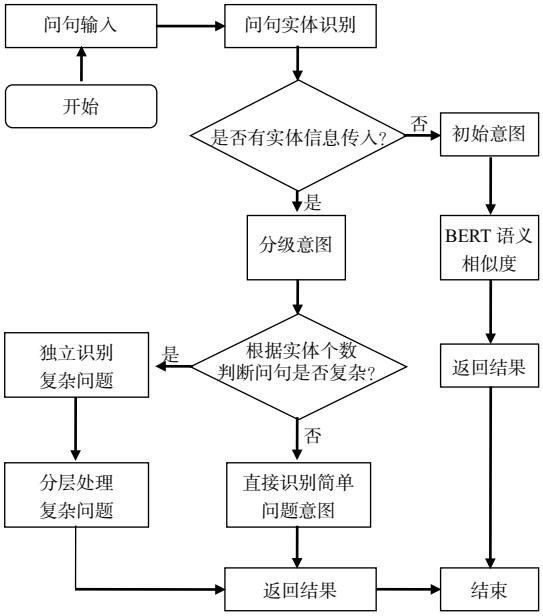

图5 意图识别流程

Fig.5 Flowchart of intent recognition

本文结合前期的知识图谱构建工作，设定了7类初始意图和18类分级意图，并根据分级意图设置了21个槽位，其中包括20个包含实体信息的槽位和1个通用槽位以适应不同复杂程度的问题处理，分别如表3和表4所示。

表3 初始意图类别

Table 3 Initial intent categories

初始意图标签	标签解释
greet	问候语
goodbye	结束语
isbot	询问是否为机器人
thank	表达对回复的感谢
disagree	表达对回复的不满意
accept	用于二次确认(肯定)
deny	用于二次确认(否定)

表4　分级意图类别

Table 4　Classification intent categories

分级意图标签	标签解释	分级意图标签	标签解释
definition	专有名词定义	publish	发布的内容
influence	某特征的影响因素	publisher	发布者信息
place	烟叶类型的产地信息	print_time	内容发布时间
eigenvalue	某分级指标的具体描述	colour	某特征的颜色信息
subordination	两项特征的从属关系	rec_method	某特征的识别方法
designation	专有名词的代号信息	rec_effect	某特征识别方法的具体描述
quantity	数量关系	stipulate	文件规定内容
variable_value	变量值信息	feature_des	表示某特征的信息
use	使用分级方法	rank_judgment	根据特征判断等级

BERT模型可用于在大量语料中学习文本和语义的特征向量表示，将学习到的特征用于下游的文本分类任务，可以大幅度提升性能。本文在初始意图模块使用了单一的BERT预训练模型进行语义相似度判别，对于初始意图这种简单的语义信息解析任务，该模型有很好的效果，而对于较为复杂的分级意图处理，该模型的效果还有待进一步提升。因此，针对使用单一预训练模型在分级意图识别上效果欠佳的问题，本文结合TextCNN在文本分类任务上的良好效果以及注意力机制在特征关注度上的优势，在原有的BERT预训练模型的基础上加入TextCNN模块和自注意力机制模块以提升模型整体效果。分级意图识别模型结构如图6所示。

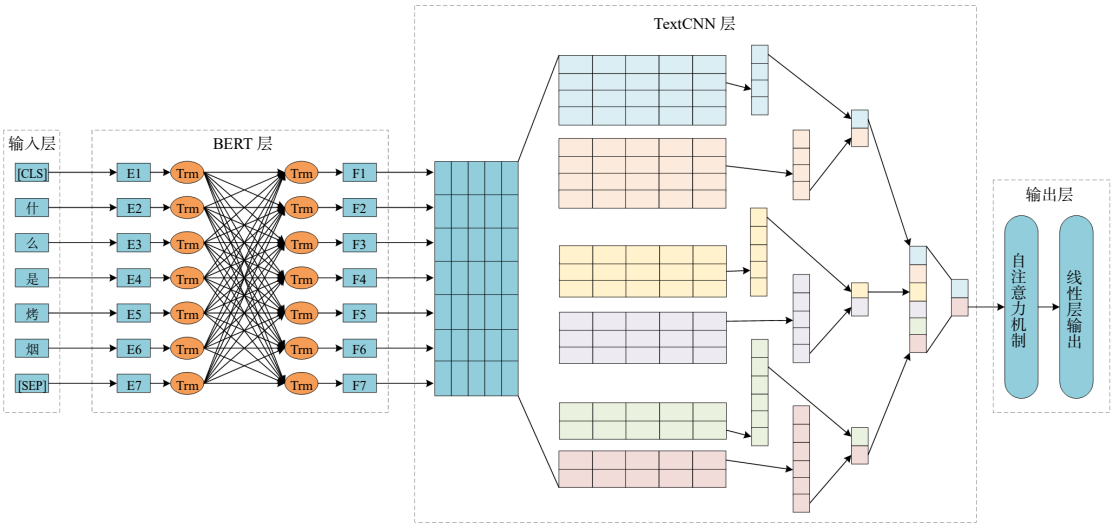

图6　分级意图识别模型结构

Fig.6　Structure of hierarchical intent recognition model

本文所设计的分级意图识别模型BERT-TextCNN-Attention主要包含文本嵌入层、BERT层、TextCNN层以及包含自注意力机制的输出层。由图6可以看出：文本数据“什么是烤烟？”经过文本嵌入层处理后，输入BERT生成向量矩阵$\boldsymbol{E}\in\mathbf{R}^{nd}$，若$\boldsymbol{E}_i$为第$i$个词的词向量，则长度为$n$的输入可表示为$\boldsymbol{E}$：

$$\boldsymbol{E}=[E_1+E_2+\cdots+E_n] \tag{17}$$

词向量在卷积核中进行卷积操作，其中卷积核大小为[2,3,4]，卷积核w与向量矩阵$\boldsymbol{E}$的第i个窗口$x_{i\sim(i+h-1)}$内的词向量进行卷积得到特征c_i。

$$c_i=f(wx_{i\sim(i+h-1)}+b) \tag{18}$$

其中，f为非线性激活函数，b为偏置项，h为卷积核的高度，卷积核w与所有向量矩阵中的词向量进行卷积后得到特征图c：

$$c=[c_1+c_2+\cdots+c_{n-h+1}] \tag{19}$$

特征图c经过文本卷积层的最大池化操作后，再将所有经过最大池化的特征进行联合得到新的特征z。

$$z=C(c_1',c_2',\cdots,c_k') \tag{20}$$

其中：$C(\cdot)$表示联合操作，z 表示不同尺寸卷积核产生的特征图。新的特征 z 需要经过自注意力机制层处理得到 $A(z)$，然后由线性层通过sigmoid函数得到模型输出$\boldsymbol{\beta}$：

$$\boldsymbol{\beta}=\sigma[A(z)\boldsymbol{W}'+\boldsymbol{b}] \tag{21}$$

其中，$A(z)$表示经注意力机制处理后的特征；σ为sigmoid函数。

最终，由示例问句“什么是烤烟？”可得到结果{0.979，0，专有名词定义}，其中意图概率值为0.979，意图类型为专有名词定义，意图类型编号为0。以上结果充分展示了模型问句意图识别效果，说明该模型可用于后续任务的处理。

3 实验

3.1 实验数据集

本文数据集主要包括2个部分：一部分用于构建烟叶分级知识图谱的数据集(见表5)；一部分用于构建问答系统的数据集(见表6)。

表5 知识图谱构建数据集

Table 5 Knowledge graph construction dataset

数据集	数据来源	实体类型	数据格式	数据量
命名实体识别数据(数据集1)	专业书籍、资料、文献(烟草工业创新平台)	见表1	BIO	标注后约40万字符
关系抽取数据(数据集2)	专业书籍、资料、文献(烟草工业创新平台)	见表2	<rel, object, subject, text>	标注后约3万条数据

表6 知识问答构建数据集

Table 6 Dataset of trivia construction

数据集	数据来源	实体/意图类型	数据格式	数据量
问句实体识别(数据集3)	问句标注结合图谱实体数据	见表1	BIO	融合后约60万字符
初始意图(数据集4)	按数据格式构造	见表3	[text, label]	约700条数据
分级意图(数据集5)	设置问句泛化模板并采用抽取到的实体替换占位符生成意图数据	见表4	[text, label, label_id]	约15 000条数据

3.2 模型评价指标

本文的命名实体识别、关系抽取、问句实体识别以及意图识别部分均采用 P (准确率)、R (召回率)和 F_1(调和平均值)进行评价。

$$P=\frac{T_{\mathrm{P}}}{F_{\mathrm{P}}+T_{\mathrm{P}}}\times 100\% \tag{22}$$

$$R=\frac{T_{\mathrm{P}}}{F_{\mathrm{N}}+T_{\mathrm{P}}}\times 100\% \tag{23}$$

$$F_1=\frac{2PR}{P+R}\times 100\% \tag{24}$$

式中：F_{P}为错误识别数，T_{P}为正确识别数，F_{N}为语料库中存在但未被识别的样本数。

3.3 实验环境

实验环境设置如表7所示。各部分实验超参数设置如表8所示。

表7 实验环境设置

Table 7 Experimental environment settings

实验环境	配置
操作系统	Windows 10
GPU配置	NVIDIA GeForce RTX 3060
显存	12 GB
开发环境	Python 3.9
开发框架	Pytorch 1.13.0

表8 各部分超参数设置

Table 8 Hyperparameter settings of each part

命名/问句实体识别超参数	值	关系抽取超参数	值	分级意图识别超参数	值
批处理大小	12	批处理大小	8	批处理大小	64
词嵌入维度	100	最大序列字符数	128	文本最大字符数	128
隐层大小	256	位置嵌入数量	60	词嵌入维度	768
标签分类数	41	位置嵌入维度	5	意图分类数	18
学习率	1×10^{-5}	学习率	2×10^{-5}	学习率	1×10^{-3}
训练轮数	50	训练轮数	30	训练轮数	100

3.4　图谱构建实验结果

3.4.1　命名实体识别

数据经doccano标注后，按照BIO格式输出并按照8∶1∶1的数量比例分为训练集、测试集和验证集，然后在数据集1中采用本文的命名实体识别模型进行解析，提取其中的分级实体。为了验证本模型的效果,将本文的BERT-BiGRU-MHSA-CRF模型与常用的知识图谱构建模型BiLSTM-CRF、BiGRU-CRF、BERT-BiLSTM-CRF、BERT-BiGRU-CRF、BERT-BiLSTM-MHSA-CRF进行对比实验，结果如表9所示。

表9　命名实体识别结果

Table 9　Results of named entity recognition

模型	P/%	R/%	F_1/%
BiLSTM-CRF	90.35	82.05	83.75
BiGRU-CRF	91.85	85.33	87.12
BERT-BiLSTM-CRF	90.61	89.38	89.60
BERT-BiGRU-CRF	91.98	88.96	90.35
BERT-BiLSTM-MHSA-CRF	90.40	91.82	91.04
BERT-BiGRU-MHSA-CRF	92.52	91.51	91.95

由表9可以看出：在相同的标注数据集上，命名实体识别模块采用BiLSTM-CRF、BiGRU-CRF、BERT-BiLSTM-CRF、BERT-BiGRU-CRF、BERT-BiLSTM-MHSA-CRF 和 BERT-BiGRU-MHSA-CRF 模型的F_1分别为83.75%、87.12%、89.60%、90.35%、91.04%、91.95%。相较于其他模型，本文模型的准确率、召回率以及F_1均有所提升，说明本文设计的BERT-BiGRU-MHSA-CRF命名实体识别模型可以有效地识别分级实体。

3.4.2　关系抽取

关系抽取模块同样采用doccano工具标注的数据进行实验，为了验证本文采用的融合了分级实体位置信息的BERT-BiLSTM-Attention模型的效果，将其与其他常用的关系抽取模型BilSTM、BERT-BilSTM和不添加实体位置信息的BERT-BiLSTM-Attention在数据集2上进行对比实验，结果如表10所示。

表10　关系抽取实验结果

Table 10　Results of relational extraction experiment

模型	P/%	R/%	F_1/%
BilSTM	83.15	82.62	83.66
BERT-BilSTM	87.50	91.30	89.36
BERT-BiLSTM-Attention (不含实体位置信息)	95.69	95.69	95.69
BERT-BiLSTM-Attention (包含实体位置信息)	97.62	97.60	97.60

由表10可见，关系抽取模块采用融合分级实体位置信息的BERT-BiLSTM-Attention模型，相较于BilSTM和BERT-BilSTM模型其准确率分别提升了14.47和10.12个百分点，召回率分别提升14.98和6.3个百分点，F_1分别提升13.94和8.24个百分点，证明本文模型在烟叶分级知识关系抽取任务上的效果优于BilSTM和BERT-BilSTM的效果；在模型中添加烟叶分级实体位置信息使F_1提升1.91个百分点，这说明向模型中添加额外的语义信息可以提升模型的分类能力。

3.5　问答系统实验结果

3.5.1　问句实体识别

问句实体识别模型采用与图谱构建相同的模型，但是在原有数据集的基础上扩充了问句实体标注数据集。为了验证本模型在烟叶分级问句方面的实体识别效果，将本文的模型同其他常用的问句实体识别模型在数据集3上做对比实验，对比模型分别为BiLSTM-CRF、BiGRU-CRF、BERT-BiLSTM-CRF、BERT-BiGRU-CRF、BERT-BiLSTM-MHSA-CRF。实验结果如表11所示。

表11　问句实体识别结果

Table 11　Results of entity recognition of questions

模型	P/%	R/%	F_1/%
BiLSTM-CRF	90.58	87.72	85.04
BiGRU-CRF	91.78	89.99	88.26
BERT-BiLSTM-CRF	91.50	89.54	90.34
BERT-BiGRU-CRF	91.63	91.16	91.34
BERT-BiLSTM-MHSA-CRF	92.42	92.62	92.66
BERT-BiGRU-MHSA-CRF	93.68	93.58	94.12

从表11可以看出，各个问句实体识别模型在现有数据集中的整体识别效果相较于图谱构建部分的命名实体识别模型均有提升，本模型在问句实体识别任务上相较于知识图谱构建实验中命名实体识别任务的准确率(表9)提升了1.16个百分点，召回率提升了2.07个百分点，F_1提升了2.17个百分点，

同时本模型的P、R、F_1这3项指标均比其他模型的高，说明本模型可以较好地处理烟叶分级领域的问句实体识别任务。

3.5.2 分级意图识别

将意图识别部分划分为初始意图和分级意图。初始意图使用BERT预训练模型进行语义相似度处理，然后输入自然语句对模型效果进行检测，发现所有回复均与预期一致，验证了该模型对初始意图识别的效果；分级意图主要采用BERT-TextCNN- Attention进行训练，为了验证本模型的有效性，将本模型在数据集5上进行消融实验并将其与意图识别常用模型进行对比实验，对比模型包括BiLSTM、BiLSTM-CRF、BiLSTM-CNN-CRF。实验结果如表12所示。

表12 分级意图识别实验结果

Table 12 Experimental results of classification intention recognition

模型	P/%	R/%	F_1/%
BiLSTM	95.23	82.33	83.69
BiLSTM-CRF	95.64	87.78	88.93
BiLSTM-CNN-CRF	96.20	92.05	93.18
BERT-TextCNN-Attention	99.16	98.63	98.77
BERT-TextCNN	95.26	94.12	94.78
TextCNN	91.83	86.99	88.52

从表12可以看出：意图识别对比模型BiLSTM、BiLSTM-CRF、BiLSTM-CNN-CRF对应的F_1分别为83.69%、88.93%、93.18%，而本模型的F_1为98.77%，表明此模型相较于其他意图识别模型在烟叶分级多类问题识别中的效果更好，可以为问答系统的性能提供保障。在消融实验中，本模型在去掉注意力机制后F_1降低了3.99个百分点，可见本文添加的注意力机制模块可以加强意图识别特征的关注度，验证了该模块的有效性；在去除注意力机制的基础上再去掉BERT预训练模块，F_1继续下降6.26个百分点，由此可见，预训练模型的使用可以大幅度地提升模型的整体效果，使分级意图的识别率提高。

3.5.3 问答系统效果测试

为了验证本文问答系统的整体性能，输入自然问句进行测试，观察回复的答案是否满足预期结果，由此可以检验问句实体识别模块和多阶段意图识别模块的有效性。问答系统部分测试结果如表13所示。

表13 问答系统测试结果

Table 13 Test results of question answering system

意图类别	自然语句	预期结果	测试结果
初始意图	你好	greet标签对应的回复模板	与预期一致
	再见	goodbye标签对应的回复模板	与预期一致
分级意图(属性)	什么叫做烤烟?	返回烤烟定义	与预期一致
分级意图(简单关系)	中部柠檬黄1级代号是什么?	返回对应的代号信息	与预期一致
	烤烟在哪些地方生产?	返回烤烟产地信息	与预期一致
分级意图(复杂关系)	具有“部位特征”的是属于什么部位?	返回部位判断	与预期一致
	具有“颜色特征”的是属于什么颜色?	返回颜色判断	
	具有“其他特征”的是属于什么等级?	结合部位和颜色返回等级判断	

由表13可以看出：无论是在初始意图还是分级意图下，自然语句的解析结果与预期结果基本一致，对于复杂关系即存在多层次处理的问题也能较好地进行识别，但烟叶分级国家标准中将烟叶等级划分为42级，因此某些等级之间的区分度不大，可能会造成本文智能问答系统在根据用户输入的特征判断等级时无法准确地进行判断。

4 结论与讨论

本文设计的烟叶分级领域问答系统充分利用了多源数据，并通过命名实体识别和关系抽取任务保证了构建图谱的质量，为智能问答系统提供了有效的数据支撑；为了使图谱数据得到有效利用，通过将图谱构建中命名实体识别部分的数据与问答数据相结合，用于问句实体识别模型的训练，提升了问句实体识别效果；同时，将意图识别部分划分为初始意图和分级意图，先通过初始意图对用户输入进行过滤，再由分级意图处理用户不同复杂程度的烟叶分级问题需求。结果显示，与现有的领域问答系统常用模型相比，引入图谱数据的问句实体识

别模块和分为多阶段、多任务的意图识别模块均取得更好的结果。虽然本文的智能问答系统取得了较好的效果，但要适应用户各种各样的问题还需扩充图谱数据以满足需求。

参考文献：

[1] 陈婷，赵晓琳，张冀武，等. 基于GA-RELM多特征优选的烟叶多部位正反面识别方法[J]. 湖南农业大学学报(自然科学版)，2025，51(1)：113–122.

[2] 闫悦，郭晓然，王铁君，等. 问答系统研究综述[J]. 计算机系统应用，2023，32(8)：1–18.

[3] HAO X J，JI Z，LI X H，et al. Construction and application of a knowledge graph[J]. Remote Sensing，2021，13(13).

[4] 刘园园，李劲华，赵俊莉. 基于语义解析的领域问答系统的设计与实现[J]. 计算机应用与软件，2021，38(11)：42–48，97.

[5] LIN J J，ZHAO Y Z，HUANG W Y，et al. Domain knowledge graph-based research progress of knowledge representation[J]. Neural Computing and Applications，2021，33(2)：681–690.

[6] 刘燕，傅智杰，李姣，等. 医学百科知识图谱构建[J]. 中华医学图书情报杂志，2018，27(6)：28–34.

[7] QIN H C，YAO Y H. Agriculture knowledge graph construction and application[J]. Journal of Physics：Conference Series，2021，1756(1)：012010.

[8] 田梦晖，陈明，席晓桃. 融合Albert模型的珍稀濒危植物知识图谱的构建[J]. 湖南农业大学学报(自然科学版)，2023，49(5)：616–623.

[9] LI J，SUN A，HAN J，et al. A survey on deep learning for named entity recognition[J]. IEEE transactions on knowledge and data engineering，2020，34(1)：50–70.

[10] WU Y R，MAO W Q，FENG J. AI for online customer service：intent recognition and slot filling based on deep learning technology[J]. Mobile Networks and Applications，2022，27(6)：2305–2317.

[11] WELD H，HUANG X，LONG S，et al. A survey of joint intent detection and slot filling models in natural language understanding[J]. ACM Computing Surveys，2022，55(8)：1–38.

[12] 于舒娟，毛新涛，张昀，等. 基于词典和字形特征的中文命名实体识别[J]. 中文信息学报，2023，37(3)：112–122.

[13] 李冬梅，罗斯斯，张小平，等. 命名实体识别方法研究综述[J]. 计算机科学与探索，2022，16(9)：1954–1968.

[14] 祁鹏年，廖雨伦，覃飙. 基于深度学习的中文命名实体识别研究综述[J]. 小型微型计算机系统，2023，44(9)：1857–1868.

[15] AN Y，XIA X Y，CHEN X L，et al. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF[J]. Artificial Intelligence in Medicine，2022，127：102282.

[16] 陈娜，孙艳秋，燕燕. 结合注意力机制的BERT-BiGRU-CRF中文电子病历命名实体识别[J]. 小型微型计算机系统，2023，44(8)：1680–1685.

[17] YANG X，XIAO Y. Named entity recognition based on BERT-MBiGRU-CRF and multi-head self-attention mechanism[C]//2022 4th International Conference on Natural Language Processing(ICNLP). Xi'an, China: IEEE，2022：178–183.

[18] WU T F，WANG M，XI Y F，et al. Intent recognition model based on sequential information and sentence features[J]. Neurocomputing，2024，566：127054.

[19] 余建明，刘赫，单连飞，等. 基于ALBERT和RE2融合模型的电网调度意图识别方法[J]. 电力系统保护与控制，2022，50(12)：144–151.

[20] 郭旭超，郝霞，姚晓闯，等. 农业病虫害知识问答意图识别与槽位填充联合模型研究[J]. 农业机械学报，2023，54(1)：205–215.

[21] NASAR Z，JAFFRY S W，MALIK M K. Named entity recognition and relation extraction：state-of-the-art[J]. ACM Computing Surveys(CSUR)，2021，54(1)：1–39.

[22] 王远强. 烟叶特征数字要素表征与分级方法研究实现[D]. 昆明：昆明理工大学，2023.

责任编辑：伍锦花

英文编辑：张承平