**硕士学位论文**

基于知识蒸馏的农业知识图谱构建研究

**RESEARCH ON THE CONSTRUCTION TECHNOLOGY OF AGRICULTURAL KNOWLEDGE GRAPH BASED ON KNOWLEDGE DISTILLATION**

**宋锦文**

**哈尔滨工业大学**

**2022年6月**

国内图书分类号：TP391.1 学校代码：10213

国际图书分类号：004.91 密级：公开

**硕士学位论文**

基于知识蒸馏的农业知识图谱构建研究

|  |  |  |
| --- | --- | --- |
| 硕士研究生 | ： | 宋锦文 |
| 导 师 | ： | 王玲教授 |
| 申请学位 | ： | 工程硕士 |
| 学科 | ： | 电子信息（计算机技术） |
| 所 在 单 位 | ： | 计算学部 |
| 答 辩 日 期 | ： | 2022年6月 |
| 授予学位单位 | ： | 哈尔滨工业大学 |

Classified Index: TP391.1

U.D.C: 004.91

Dissertation for the Master Degree

**RESEARCH ON THE CONSTRUCTION TECHNOLOGY OF AGRICULTURAL KNOWLEDGE GRAPH BASED ON KNOWLEDGE DISTILLATION**

|  |  |
| --- | --- |
| **Candidate：** | Song Jinwen |
| **Supervisor：** | Prof.Wang Ling |
| **Academic Degree Applied for：** | Master of Engineering |
| **Speciality：** | Computer Technology |
| **Affiliation：** | School of Computer Science |
| **Date of Defence：** | June, 2020 |
| **Degree-Conferring-Institution：** | Harbin Institute of Technology |

# 摘 要

粮食安全问题是关系国计民生的根本性问题，病虫害、自然灾害等对我国粮食产量有着巨大的影响。在应对这些灾害时，网上与之相关的信息虽然繁多，但是存在内容表达差异以及知识零散等问题，数据存储结构以及表示方式等也各不相同，处于一种相对混乱的状态，人们难以高效地利用好这些信息。近十年里，知识图谱技术发展迅速，其能有效地描述各类事物，以及提高搜索查询的速度与质量，因此利用知识图谱相关技术来整合农业领域相关的数据，能帮助我们更快更好地应对作物生长过程中可能遇到的各类问题。

本文构建了农业知识图谱系统，实现了实体识别、关系抽取以及知识图谱中实体和关系的查询等功能。在构建知识图谱的过程中，首先根据《中国农业病虫害》、百度百科以及各类农业相关的数据，构建了农业领域本体。然后采用基于知识蒸馏的BiLSTM-CRF命名实体识别模型以及PCNN关系抽取模型等方法，对结构化和非结构化的数据进行知识抽取，并将抽取出来的知识存储在neo4j图数据库中，共计13983个实体和26498个关系。

在构建知识图谱过程中，往往需要挖掘非结构化的数据，而农业领域现存公开的带标签的数据集稀少，因此针对带标签的数据少的情况，本文提出了一个基于知识蒸馏的迁移学习模型，首先训练一个领域鉴别器，其能够对源域和目标域的数据有一个较为准确的分类，然后通过领域鉴别器选出少量目标域的数据，之后仅使用该部分数据，运用知识蒸馏相关的技术，能提升目标域模型的效果。通过实验结果发现，对于一个领域的数据，我们只需标注不到十分之一的数据，相比只用这部分数据进行训练，通过知识蒸馏后能提高模型的效果，因为领域鉴别器里能选出源域对目标域有效的那部分数据。通过这种方式，在大幅缩减数据标注所耗费的人力物力的情况下，模型能达到一个不错的效果。

关键词:命名实体识别；关系抽取；知识图谱推理；知识蒸馏

# Abstract

Food security is a fundamental issue related to the national economy and the people's livelihood. Plant diseases, pests and natural disasters have a great impact on China's grain output. In response to these disasters, although there are many information related to them on the Internet, there are many problems such as differences in content expression and scattered knowledge. The data storage structure and presentation methods on the Internet are also different and still in a chaotic state. It is difficult for people to make efficient use of this information. In the past decade, knowledge graph technology has developed rapidly. It can effectively describe all kinds of things and improve the speed and quality of search and query. Therefore, using knowledge graph related technology to integrate relevant data in the agricultural field can help us deal with all kinds of problems that may be encountered in the process of crop growth faster and better.

This paper constructs an agricultural knowledge graph system, which realizes the functions of entity recognition, relationship extraction and the query of entity and relationship in the knowledge graph. In the process of constructing the knowledge graph, the agricultural domain ontology is first built according to China agricultural diseases and pests, Baidu Encyclopedia and various agricultural related data. Then, the BiLSTM-CRF named entity recognition model based on knowledge distillation and PCNN relationship extraction model are used to extract knowledge from structured and unstructured data, and the extracted knowledge is stored in the neo4j graph database, with a total of 13983 entities and 26498 relationships.

In the process of constructing the knowledge graph, it is often necessary to mine unstructured data. However, there are few publicly available labeled data sets in the agricultural field. Therefore, in view of the lack of labeled data, this paper proposes a transfer learning model based on knowledge distillation. First, a domain discriminator is trained, which can accurately classify the data in the source domain and the target domain. Then a small amount in target domain datasets is selected through the domain discriminator. By using this part of data for training, the effect of the target domain model can be improved by using knowledge distillation related technologies. Through the experimental results, we can found that for the data in a field, we only need to label less than one tenth of the data. Compared with only using this part of data for training, the model effect can be improved after knowledge distillation, because the field discriminator can select the part of data that the source domain is effective for the target domain. In this way, the model can still achieve a good effect when the human and material resources consumed by data annotation are greatly reduced.

**Keywords**：named entity recognition, relation extraction, knowledge graph reasoning, knowledge distillation

目 录

[摘 要 I](#_Toc19924)

[Abstract II](#_Toc20903)

[第1章 绪论 1](#_Toc6805)

[1.1 课题背景及研究的目的和意义 1](#_Toc25867)

[1.2 国内外研究现状 1](#_Toc28318)

[1.2.1 命名实体识别技术 1](#_Toc19702)

[1.2.2 关系抽取技术 2](#_Toc24296)

[1.2.3 知识图谱推理技术 2](#_Toc1939)

1.3 本文的主要研究内容 .3

[1.4 论文的组织结构](#_Toc447) 4

[第2章 农业知识图谱构建相关技术](#_Toc18334) 6

2.1 引言 6

[2.2 文本嵌入表示](#_Toc26126) 6

[2.3 实体识别及关系抽取技术](#_Toc21259) 6

[2.3.1 循环神经网络](#_Toc10712) 6

[2.3.2 长短期记忆网络 7](#_Toc10256)

[2.3.3 条件随机场](#_Toc20929) 7

[2.3.4 注意力机制](#_Toc8981) 8

[2.4 知识蒸馏技术](#_Toc20882) 8

[2.4.1 互信息知识蒸馏](#_Toc30112) 8

[2.5 知识图谱推理技术](#_Toc30232) 9

[2.5.1 Tucker 分解](#_Toc32061) 10

[2.5.2 SimplE模型](#_Toc17126) 10

[2.5.3 DisMult模型](#_Toc28288) 10

[2.6 本章小结](#_Toc32315) 12

[第3章 农业知识图谱模式层构建](#_Toc5988) 12

3.1 引言 12

[3.2 主要参考数据源 1](#_Toc7181)2

[3.3 农业知识图谱本体构建 1](#_Toc25103)2

[3.3.1 本体模型设计 1](#_Toc1732)2

[3.3.2 农业实体及其属性定义](#_Toc7274) 13

[3.3.3 农业语义关系定义](#_Toc22618) 15

[3.3.4 本体评估](#_Toc21840) 17

[3.4 本章小结](#_Toc5579) 17

[第4章 农业知识图谱数据层构建](#_Toc4407) 18

4.1 引言 18

[4.2 农业知识图谱构建流程](#_Toc32654) 18

[4.3 农业数据标注工具](#_Toc3366) 19

[4.4 命名实体识别](#_Toc11064) 21

[4.4.1 自注意力机制下的对抗迁移学习模型](#_Toc22263) 21

[4.4.2 基于知识蒸馏的迁移学习模型 2](#_Toc12402)3

[4.4.3 实验结果及分析](#_Toc11768) 24

[4.5 关系抽取](#_Toc16503) 29

[4.5.1 PCNN模型](#_Toc11515) 29

[4.5.2 实验结果及分析](#_Toc27610) 29

[4.6 知识图谱存储方法](#_Toc4266) 30

[4.7 本章小结](#_Toc713) 32

[第5章 农业知识图谱系统的设计与实现](#_Toc11931) 33

5.1 引言 33

[5.2 系统逻辑架构 3](#_Toc7190)3

[5.3 知识图谱推理](#_Toc13308) 34

[5.3.1 TuckER模型](#_Toc29110) 34

[5.3.2 实验结果及分析](#_Toc29941) 35

[5.4 系统主要功能模块的实现](#_Toc4064) 37

[5.4.1 实体识别及关系抽取功能](#_Toc858) 37

[5.4.2 知识图谱查询功能](#_Toc1450) 39

5.4.3 知识图谱问答功能 39

[5.5 本章小结](#_Toc11315) 40

[结 论](#_Toc9501) 41

[参考文献](#_Toc22011) 42

[攻读硕士学位期间取得创新性成果](#_Toc3324) 46

哈尔滨工业大学学位论文原创性声明和使用权限 47

[致 谢](#_Toc4503) 48

# 第1章 绪 论

## 1.1 课题背景及研究的目的和意义

“民以食为天”，粮食安全乃人类生存之要,国泰民安之基，然而病虫害、自然灾害等对我国粮食产量有着巨大的影响。在应对这些灾害时，网上与之相关的信息虽然繁多，但是存在一个物品有多种表示以及知识零散等问题，数据存储结构以及表示方式等也各不相同，处于一种相对混乱的状态，人们难以高效地利用好这些信息。近十年里，知识图谱技术发展迅速，其能有效地描述各类事物，以及提高搜索查询的速度与质量，因此利用知识图谱相关技术来整合农业领域相关的数据，能帮助我们更快更好地应对作物生长过程中可能遇到的各类问题。

由于网上的信息大多都是非结构化的数据，且大多以文字的形式进行描述，因此在构建知识图谱过程中，需要进行命名实体识别任务和关系抽取任务。近年来，随着人工智能的发展，深度学习的方法在命名实体识别任务和关系抽取任务中取得了非常好的成果，然而深度学习模型通常都需要大量标注数据作为支撑，当其在标注数据匮乏的领域应用时，往往因为会训练数据不足导致模型效果不够理想。针对这一问题，采用迁移学习的方法能较好的进行处理，其主要思想是先将模型在数据丰富的领域进行训练，得到一个预训练模型，再将该模型在数据匮乏的领域进行微调，让其适应目标域上的学习任务。因此，采用迁移学习的方法可以较好的解决数据匮乏的问题，还可以利用源域中丰富的数据资源来提高目标域任务中模型的效果。

## 1.2 国内外研究现状

### 1.2.1 命名实体识别技术

命名实体识别（Named Entity Recognition , NER）是识别出文本中的命名实体，是自然语言处理领域中的一类基础的任务，许多自然语言处理相关的任务都需要命名实体识别任务的支持，比如知识图谱推理，问答系统等。

早期命名实体识别任务主要是基于字典和规则进行的。基于规则的NER系统主要有NetOwl[1]，Facile[2]以及FASTU[3]等，这些系统通常需要大量人力物力来构造语义和语法规则，且只能针对特定的文本环境，一旦文本环境发生巨大变化，识别效果往往会变差。

随着传统机器学习技术的发展，不少传统机器学习的方法被用在命名实体识别任务中，主要包括隐马尔科夫模型（Hidden Markov Model）[4]、最大熵模型（Maximum Entropy）[5]、支持向量机（Support Vector Machine）[6]、决策树模型（Decision Tree）[7]以及条件随机场（Conditional Random Field, CRF）[8]等。

近年来，深度学习的方法在命名实体识别中逐渐成为主导，包括卷积神经网络（Convolutional Neural Network）[9]、循环神经网络（Recurrent Neural Network）[10]、长短期记忆网络（Long Short Term Memory Network）[11]以及门控循环单元（Gated Recurrent Unit）[12]等。 Huang等人[13]将双向长短期记忆模型BiLSTM和条件随机场CRF结合起来，并且在之后许多命名实体识别的研究中，都是在此结构基础上进行改造的。

针对目标域训练数据匮乏的问题，Canasai等人[14]将多任务和预训练结合起来通过易获得的句子级标签来提高模型在低资源领域下的效果。[Kevin](https://dblp.org/pid/78/6661.html)等人[15]提出交叉视野训练的方法，利用了大量无标注的文本数据提高了模型的效果。Marek[16]通过学习预测序列中每个词的周围的词来提高模型效果。Liu等人[17]通过多任务学习和混合专家系统的方法，在不利用目标域数据的情况下提高了模型的效果。Bari等人[18]通过字级对抗和参数共享以及特征增强的方式来提高模型效果。Pengfei等人[19]采用领域对抗的方式来提高模型效果。

### 1.2.2 关系抽取技术

关系抽取（relation extract）是识别出文本中两个实体之间的关系，同样是自然语言处理领域中的一类基础的任务。

早期关系抽取任务主要是基于规则进行的。Aone等人[20]通过人工编写关系模板的方式进行关系抽取任务。Fukumoto等人[21]通过两个实体间的谓词来分析实体间的关系。基于规则的方式通常需要消耗大量的人力物力来构建规则，且适用于特定的文本环境。

随着传统机器学习技术的发展，Kambhatla等人[22]通过最大熵模型来实现关系抽取任务。Miao等人[23]则通过CRF来实现关系抽取任务。之后，深度学习技术逐渐成为主流，Zeng等人[24]通过将句子根据实体位置进行分段，提出了PCNN模型，进行关系抽取。Qin等人[25]采用对抗学习的方法来提升模型效果。Sun等人[26]采用强化学习的方法来识别带噪音的数据。

### 1.2.3 知识图谱推理技术

知识图谱推理技术是基于知识图谱中已有的实体和关系来推理出未知的实体和关系，知识图谱推理技术也可以用作知识图谱补全，因其可以用来推理出知识图谱中缺失的三元组，达到补全知识图谱的效果。

目前静态知识图谱推理算法以表示学习为主，可分为以下几类：

（1）基于距离转移的模型。该方法是根据实体和关系在映射空间中的距离来判定实体及关系之间的联系。该类模型最经典的是Borders等人[27]提出的TransE模型，该模型采用欧式距离，将关系看作是头实体和尾实体之间的距离向量，即头实体的向量表示加伤关系的向量表示接近尾实体的向量表示，但该模型缺很难处理自反、多对多等复杂的关系。在此之后Wang等人[28]在transE的基础上引入了超平面，提出了tranH模型。Lin等人[29]在transE的基础上构造了多个欧式空间，提出了transR模型。Ji等人[30]提出的transD模型则是对tranR进行了优化。

（2）基于语义匹配的模型。该方法是通过实体及关系之间的语义相似度来判定实体及关系之间的联系。该类模型的代表是Nickel等人[31]提出的RESCAL模型，该模型用向量表示实体，用矩阵表示关系，通过张量分解的方法来表示实体和关系之间的联系。Yang等人[32]在RESCAL模型的基础上将关系矩阵限制为对角阵，提出了DistMult模型。Ivana等人[33]则是通过Tucker分解的方式，提出了TuckER模型。

（3）基于神经网络的模型。该方法是通过神经网络来学习实体及关系之间的联系。Michael等人[34]通过图卷积神经网络来学习实体之间的领域特征，提出了R-GCN模型。Pasquale等人[35]则将卷积神经网络用于知识图谱推理。

## 1.3 本文的主要研究内容

本文主要研究内容如图1-1所示，主要包括以下三个方面：

（1）农业知识图谱的模式层构建。知识图谱是结构化的语义知识库，能用来描述各类事物之间的关系，事物之间的关系概念模型构成了知识图谱的模式层，即本体模型。知识图谱的模式层构建是基于领域的知识特点以及应用属性，对事物进行抽象和建模，主要是实体定义、关系定义以及属性定义。本文基于《中国农业病虫害》和农业领域相关专家的知识以及农业相关的数据资源，研究构建了农业知识图谱的本体模型。

（2）农业知识图谱的数据层构建。在完成了农业知识图谱模式层的构建后，本文基于农业知识图谱模式层构建的本体模型，对结构化和非结构化的数据利用命名实体识别和关系抽取相关的技术进行抽取，从中抽取出符合本体构建的知识，考虑标注工作的时间和金钱的成本，可能导致标注的数据不足，本次研究将引入迁移学习以及知识蒸馏等技术，提高模型对未标注部分的识别效果。最后将挖掘的数据存储在图数据库中，完成了农业知识图谱的数据层构建。

（3）农业知识图谱系统的设计与实现。在完成农业知识图谱模式层和农业知识图谱数据层的构建工作后，依据模式层构建的本体以及数据层提供的三元组数据，完成了农业知识图谱构建工作，本文在其农业知识图谱的基础上设计并实现了农业知识图谱系统，其能让实体识别和关系抽取等技术更加直观，也能让用户通过查询功能或者问答功能来快速获得知识图谱中的信息。

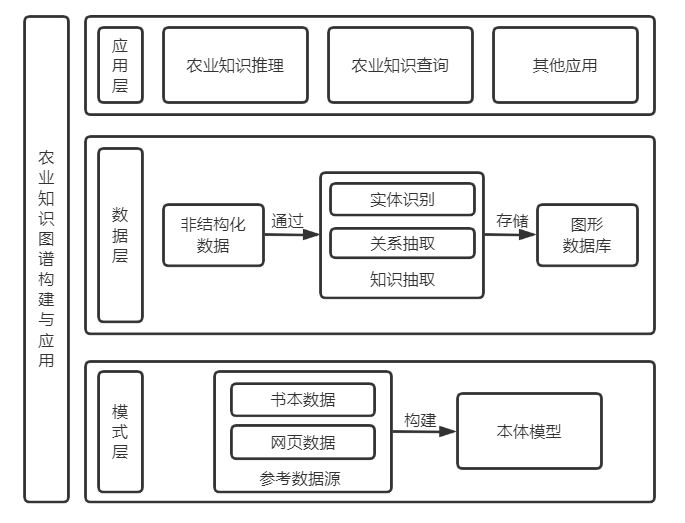


图1-1 本文主要研究内容

## 1.4 论文的组织结构

本文共分为五章，各章的组织结构如下：

（1）第一章：绪论。本章首先介绍研究背景、研究目的以及研究意义，随后对构建知识图谱所用到的技术的研究现状进行介绍，最后介绍了本文的主要研究内容。

（2）第二章：农业知识图谱构建相关技术。本章主要介绍后续章节中农业知识图谱构建所用到的相关技术与理论。

（3）第三章：农业知识图谱的模式层构建。本章主要介绍农业知识图谱本体模型，并对其中定义的实体以及关系进行介绍。

（4）第四章：农业知识图谱的数据层构建。本章主要介绍进行农业数据挖掘所用到的一些方法及其实验结果，并对本文提出的基于知识蒸馏的模型进行介绍，最后介绍这些数据的存储方法。

（5）第五章：农业知识图谱系统的设计与实现。本章主要对基于农业知识图谱搭建的系统进行介绍，首先介绍农业知识图谱系统的逻辑架构，随后介绍知识推理的方法及其实验结果，最后对系统实现的功能进行展示。

# 第2章 农业知识图谱构建相关技术

## 2.1 引言

本章将从以下几个方面介绍了后续章节中农业知识图谱构建所用到的相关技术与理论：首先在第二小节介绍了文本嵌入表示，然后在第三小节介绍了进行实体识别任务和关系抽取任务所用到的相关技术，接着在第四小节介绍了知识蒸馏的相关技术，最后在第五小节介绍了知识图谱推理模型以及一些分解公式。

## 2.2 文本嵌入表示

在深度学习技术中，嵌入表示是一项重要的技术，其可以通过一个向量的方式来表达一个对象，使这个对象能够在计算机内被使用，这个对象可以是文本以及图像等。在文本中，我们可以通过向量来对一个字或者词进行嵌入表示，由于文本中的字或者词被表示成了向量的形式，不同字或者词之间的相似度可以通过KL散度、欧式距离等方式描述。在嵌入表示中，word2vec是一项重要的工作，Word2vec 可以根据给定的语料库，通过优化后的训练模型快速有效地将一个词语表达成向量形式。word2vec有两种训练方式，分别是连续词袋（CBOW）和skip-gram。

## 2.3 实体识别及关系抽取技术

### 2.3.1 循环神经网络

在使用深度学习处理时序问题时，循环神经网络（Recurrent Neural Network, RNN）是最常使用的模型之一。由于RNN在t时间片时会将t-1时间片的隐节点作为当前时间片的输入，使其能获得之前时间片的信息，并用于当前时间片的计算，从而RNN在时序数据上有着很好的表现，其中RNN的结构如图2-1所示。RNN的数学表达式如公式(2-1)所示。其中ht为t时刻的隐节点，w为对应的权重，xt为t时刻的输入，σ为激活函数。

(2-1)

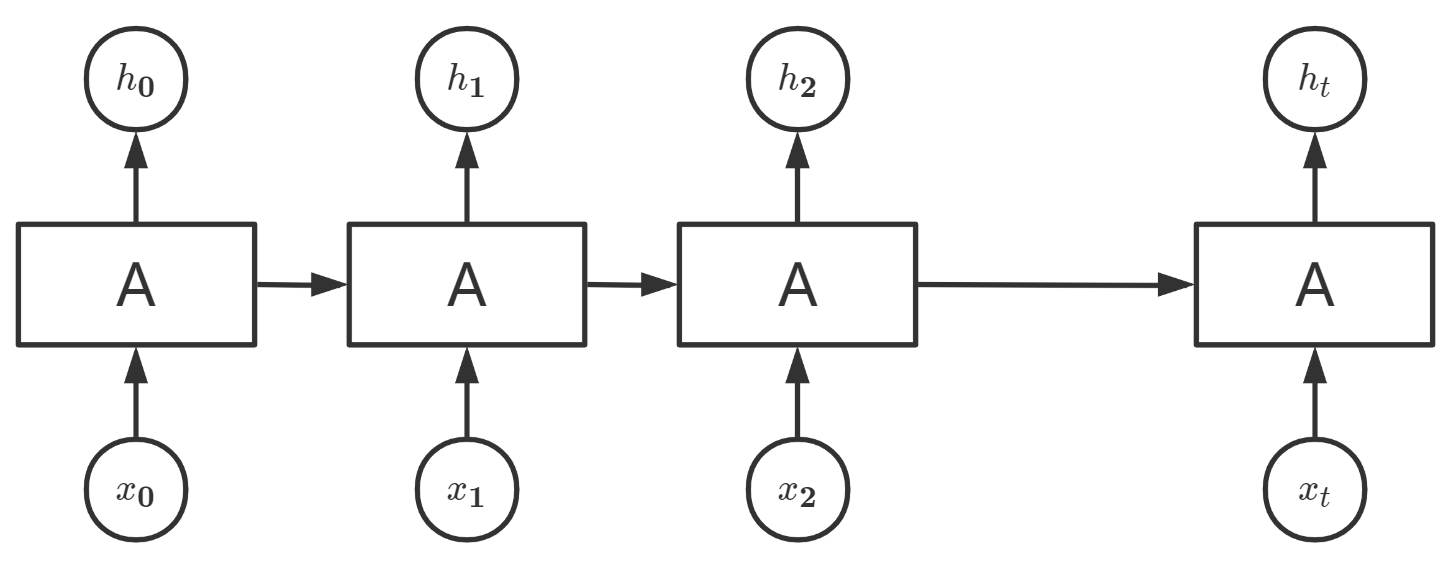


图2-1 RNN网络结构

### 

### 2.3.2 长短期记忆网络

在RNN网络中，当神经网络的节点经过许多阶段的计算后，离当前时刻比较远的时刻的信息对当前时刻的影响可能会越来越小甚至消失。相比于RNN，长短期记忆网络(Long Short Term Memory Network, LSTM)引入了门控机制用于控制特征的流通和损失，从而使其能够处理长期依赖问题。

LSTM引入了3个门控单元，分别用于控制信息的存储，更新和丢弃。细胞状态从上一时刻到当前时刻的转移通过输入门和遗忘门共同控制，输入门决定当前时刻有多少信息被吸收，遗忘门决定上一时刻的细胞状态有多少信息不被遗忘。LSTM通过门控单元以及细胞状态形成循环，给梯度的长距离传播提供了路径，改变了之前循环神经网络RNN中信息和梯度的传播方式，从而解决了长期依赖问题。

### 2.3.3 条件随机场

条件随机场（Conditional Random Field, CRF）是对多个变量在给定观测值后的条件概率进行建模。若令x = {x1, x2, ... , xn}为观测序列，y = {y1, y2, ... ,yn}为与之相对应的标记序列，则条件随机场的目标是构建条件概率模型P(y|x)。

在网络模型中加入CRF层可以向最终的预测标签添加一些约束，可以确保它们是合理的，这些约束可以由CRF层在训练过程中从训练数据集中学习。在命名实体识别任务中，假如采用BIO的标注方式，约束条件规定了实体的第一个字一定是以“B-”开头的标签，在标签“... B-label1 I-label2 I-label3 ...”中，label1、label2和label3应该是相同的实体标签等。

在CRF层的损失函数中，有两种类型的分数。第一个是emission分数，这些emission分数来自前一层网络的标签分类结果。第二个是transition分数，表示从一个标签到另一个标签的得分，比如在一个句子中，标签“B-Person”后面往往接的是“I-Person”或者“O”，而不会接“I-”开头的其他类型的标签。通过CRF层后，会使得原分类结果更加合理，效果也更加好。

### 2.3.4 注意力机制

注意力机制源于对人类视觉的研究，其可以在众多的输入信息中聚焦于对当前任务更为关键的信息，降低对其他信息的关注度，甚至过滤掉无关信息，从而提高任务处理的效率和准确性。在自然语言处理领域中，句子中的每个字都对应着三个向量：Query向量，Key向量和Value向量。每个字的Query向量与其他字的Key向量相乘的结果，表示该字与其他字的一个关联度，可以用来调整Value向量，其公式表示如公式(2-2)所示。通常情况下，Key向量的值与Value向量的值相同。当Query向量、Key向量和Value向量的值都相同时，称为自注意力机制(Self-Attention)。

(2-2)

## 2.4 知识蒸馏技术

知识蒸馏（knowledge distillation, KD）是模型压缩的一种常用的方法，不同于模型压缩中的剪枝和量化，知识蒸馏方法则是利用性能更好的参数量更大的模型的信息来训练一个参数量小的模型，使参数量小的模型的性能更加好。这个参数量大的模型我们称之为教师模型（teacher model），参数量小的模型我们称之为学生模型（student model）。教师模型传递给学生模型的信息我们称之为知识（knowledge），而学生模型学习来自教师模型的知识的过程称之为知识蒸馏（knowledge Distillation）。

### 2.4.1 互信息知识蒸馏

互信息知识蒸馏（Variational Information Distillation, VID）方法是将知识蒸馏的最优性能定义为最大化教师模型和学生模型之间的互信息。首先互信息的定义如公式(2-3)所示。

(2-3)

互信息为教师模型的熵值减去已知学生模型的条件下的教师模型的熵值。当学生模型已知时，如果其能够使得教师模型的熵很小，这说明学生模型已经获得了能够恢复教师模型所需要的“压缩”知识，间接说明了此时学生模型已经学习的很好了，而这种情况下就说明公式(2-3)中的很小，而是已知固定的值，从而就会很大，即教师模型和学生模型之间的互信息就会很大。

如公式(2-4)所示，由于难以计算，这里利用了可变高斯来模拟，式中的大于等于操作用到了KL散度的非负性，由于蒸馏过程中和需要学习的学生模型参数无关，因此最大化互信息就转换为最大化可变高斯分布的问题。

(2-4)

如公式(2-5)所示，可以通过建立一个均值和方差可学习的高斯分布来模拟上述的。

(2-5)

## 2.5 知识图谱推理技术

设表示知识图谱中的所有实体的集合，表示知识图谱中的所有关系的集合，则知识图谱中的三元组可以表示为，其中分别表示知识图谱中的头实体和尾实体，表示这两实体间的关系。

知识图谱推理算法通常能用来补充知识图谱中缺失的三元组，使知识图谱更加完整。知识图谱推理算法可以被抽象成一个链接预测问题，即预测三元组中缺失的部分，包括根据头实体和关系来预测尾实体，根据头实体和尾实体来预测两实体间的关系以及根据关系和尾实体来预测头实体。因此，知识图谱推理技术有时也被用来进行知识图谱补全任务。

在知识图谱推理算法中，通常是以学习一个得分函数为目标的，该得分函数能够给三元组进行打分，得到的分数判断该三元组是否为真，最终的目的是能够正确评分所有确实的三元组。评分函数通常是线性模型中张量因子分解的一种特定形式，或者是非线性模型中更复杂的神经网络结构。通常，一个特定三元组的正分数表示模型预测的真实事实，而负分数表示错误的事实。

2.5.1 Tucker 分解

Tucker分解（Tucker Decomposition）是高阶的主成分分析的一种形式，它将一个张量分解成一个更小的核张量与每一维矩阵的乘积。在三模态的情况下，给定一个张量，Tucker分解输出一个张量和三个因子矩阵，，。如公式(2-6)所示，其中表示张量沿着第n维的乘积，因子矩阵A，B，C（通常是正交的）可以作为每一个维度上的主成分，核张量Z表示每一维成分之间的联系程度，通常情况下，P、Q、R分别比I、J、K要小，所以Z可以被认为是X的压缩版。Tucker分解通常不是唯一的。当核张量Z是对角化的且P = Q = R时，CP分解便是Tucker分解的一种特殊形式。

(2-6)

### 2.5.2 SimplE模型

在canoncial Polyadic（CP）分解中，每个实体e由两个向量进行嵌入表示，每个关系r由一个向量进行嵌入表示。表示实体e作为头实体时的特征，表示实体e作为尾实体时的特征。于是对一个三元组，其得分表示为，其中表示向量点积，因此对一个实体e的两种embedding表示和是相互独立的，即在以作为三元组进行训练时，只会更新和，不会更新和。

SimplE模型为每个实体e的嵌入表示定义了两个向量，为每个关系r的嵌入表示定义了两个向量。于是SimplE的得分函数就定义为，即三元组和的得分的平均值。SimplE模型引入反转关系来使得每个实体的两个向量表示彼此有了联系。

2.5.3 DistMult模型

在DistMult模型中，实体表示为，，其中和分别表示实体和通过one-hot的向量，W为参数方程，可以是线性或者非线性函数。关系表示为 和。其中和是特定于关系的参数，得分函数为双线性打分函数。DistMult模型则将限制为对角阵。

## 2.6 本章小结

本章主要介绍了后续章节所需要用到的相关技术与理论，首先介绍了文本嵌入表示，然后介绍了实体识别和关系抽取所用到的相关技术，包括CNN、LSTM等，接着介绍了知识蒸馏的相关技术，最后介绍了知识图谱推理的一些分解公式。

# 第3章 农业知识图谱模式层构建

## 3.1 引言

## 知识图谱是结构化的语义知识库，能用来描述各类事物之间的关系，事物之间的关系概念模型构成了知识图谱的模式层，即本体模型。知识图谱的模式层构建是基于领域的知识特点以及应用属性，对事物进行抽象和建模，主要是实体定义、关系定义以及属性定义。

## 本章将从以下几个方面介绍农业知识图谱模式层构建工作：首先在第二小节介绍了本体设计主要参考的数据及其特点；然后介绍了本体模型设计，即本体中定义的实体及其属性以及实体间的关系类型；最后简单介绍了本体评估。

## 3.2 主要参考数据源

在进行本体设计前，需要确定本体描述的专业领域并进行数据源的搜集及确定。本文所要构建的本体属于农业领域，主要参考的数据源有：（1）《中国农作物病虫害（第三版）》。全书共计24个单元，涉及1500余种农业病虫害，其中病害775种，虫害739种，重点介绍了病害的分布和危害、症状、病原、防治技术等，虫害的分布和危害、形态特征、生活习性、防治技术等。（2）百度百科数据。百度百科词条主要介绍了植物、病虫害、农药、化肥等的一些基本属性，其中植物包括别名、种属科目纲门界、形态特征、分布范围、价值等；病害包括症状、病原、防治方法等；虫害包括病虫特征、防治方法等；农药化肥则包括组成成分等。（3）《农业科学叙词表》。它是当今世界上具有较新的结构、较大的收词量的农业叙词表，总共包含64638条叙词，是一部具有科学性和规范性的农业信息检索书。（4）中国作物种质信息网。该网站共收录了340多种作物、47万份种质的信息。可以在作物病虫害知识中可以找到水稻、麦类、豆类等一些作物的病虫害信息，病害主要有症状、病原、传播途径和发病条件、防治方法等信息，虫害主要有寄主、为害特点、形态特征、生活习性和防治方法等信息。

## 3.3 农业知识图谱本体构建

### 3.3.1 本体模型设计

本体模型是在抽象层面描述由概念及其之间的关系建立而来的逻辑模型，可以直观的展示本体及其属性以及本体间的关系。本文设计的农业领域本体模型的概要图如图3-1所示。该图描述了所定义的七类实体（农作物、病害、虫害、农药、肥料、症状和自然环境）以及这些实体之间的一些关系。实体之间可能不会有关系，如农药和肥料之间就没有定义关系；也有可能存在着多种关系，如自然环境可能会促进农作物的生长，也有可能会抑制农作物的生长；还有可能是实体自身与自身存在关系，比如同类农作物之间可能存在上下位关系，虫害与虫害之间可能存在防治关系。

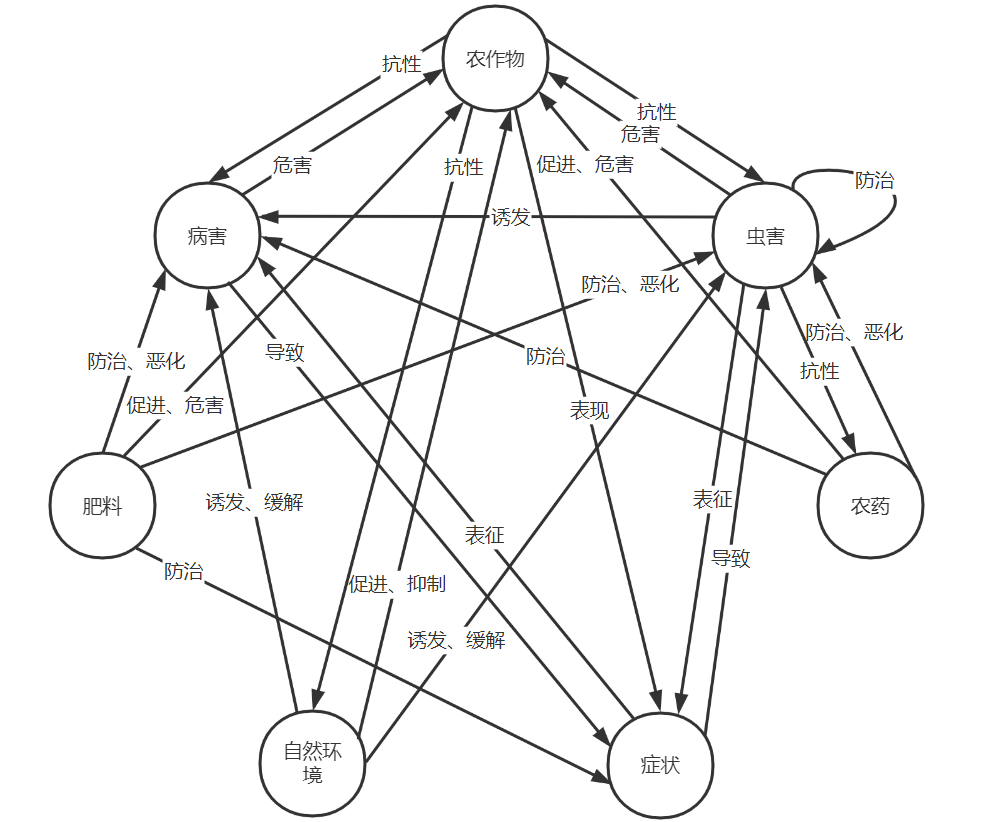


图3-1 农业领域本体模型的概要图

### 3.3.2 农业实体及其属性定义

本体设计共定义了七类实体类型，如表3-1所示，分别是农作物、病害、虫害、农药、肥料、症状和自然环境。其中，农作物可细分为大田作物、水果、蔬菜三类实体，其属性包括别称、拉丁文名称、界、门、纲、目、科、属、种、分布范围、繁殖方式、加工产品、栽培技术、价值，其中价值还可细分为食用价值、饲用价值、经济价值、加工价值、药用价值、营养价值、保健价值等。病害可根据病害成因细分为真菌病害、细菌病害、病毒病害、线虫病害、其他病害五类实体，其中其他病害是指由缺少微量元素、不良自然环境、药害、肥害等引起的病害统称，病害的属性包括别称、越冬、防治方法，其中防治方法可细分为农业防治、生物防治、物理防治、化学防治。虫害可根据病虫的类型分为昆虫纲虫害、蛛形纲虫害和其他虫害三类实体，虫害的属性包括别称、越冬、病虫特征、防治方法，其中防治方法可细分为农业防治、生物防治、物理防治、化学防治。农药根据农药的作用可细分为杀虫剂、杀螨剂、灭鼠剂、杀菌剂、除草剂、增效剂、植物生长调节剂七类实体，其属性包括农药成分、农药浓度、农药用量和使用方法。肥料根据所含元素的不同，可细分为氮肥、磷肥、钾肥、复合肥、微量元素肥五类实体，其属性包括使用方法和肥料用量。自然环境可细分为土壤温度、土壤湿度、土壤盐度、土壤酸碱度、土壤营养元素、土壤类型、空气温度、空气湿度、空气二氧化碳浓度、空气氨气浓度、风速、阳光、水。

表3-1 农业实体及其属性

|  |  |  |  |
| --- | --- | --- | --- |
| 农业实体 | 子实体 | 属性 | 定义 |
| 农作物 | 大田作物、水果、蔬菜 | 别称、拉丁文名称、界、门、纲、目、科、属、种、分布范围、繁殖方式、加工产品、栽培技术、价值 | 农业上经大片田地栽培获得的粮食类、经济类植物统称为农作物 |
| 病害 | 真菌病害、细菌病害、病毒病害、线虫病害、其他病害 | 别称、越冬、防治方法 | 由细菌、真菌、病毒、环境等引起植物发育不良、枯萎或死亡统称为病害 |
| 虫害 | 昆虫纲虫害、蛛形纲虫害、其他虫害 | 别称、越冬、病虫特征、防治方法 | 对植物生长造成影响的害虫 |
| 农药 | 杀虫剂、杀螨剂、灭鼠剂、杀菌剂、除草剂、增效剂、植物生长调节剂 | 农药成分、农药浓度、农药用量和使用方法 | 指保障、促进植物和农作物的成长所施用的杀虫、杀菌等的一类药物 |

表3-1（续表）

|  |  |  |  |
| --- | --- | --- | --- |
| 农业实体 | 子实体 | 属性 | 定义 |
| 肥料 | 氮肥、磷肥、钾肥、复合肥、微量元素肥 | 使用方法和肥料用量 | 能供给作物生长发育所需养分，改善土壤性状，提高作物产量和品质的物质 |
| 症状 | - | - | 泛指由病害、虫害、营养不足、不良环境导致的植物生理、组织结构和形态上所发生的病变特征 |
| 自然环境 | 土壤温度、土壤湿度、土壤盐度、土壤酸碱度、土壤营养元素、土壤类型、空气温度、空气湿度、空气二氧化碳浓度、空气氨气浓度、风速、阳光、水 | - | 在农田中，由水土、风、光、地域等自然事物所形成的环境 |

### 3.3.3 农业语义关系定义

本体设计共定义了十二类语义关系类型，包括抗性、危害、表现、促进、抑制、导致、表征、诱发、防治、缓解、恶化、上下位。如表3-2所示，其中第一列表示头实体类型，第一行表示尾实体类型，表中间部分对应的内容表示头实体对尾实体的关系。其中抗性关系包括农作物对病害、虫害、自然环境的抗性关系以及虫害对农药的抗性关系，指农作物的某些品种会对这些病害、虫害以及自然环境有一定的抵抗能力以及病虫的抗药性。危害关系包括病害、虫害、农药、肥料对农作物的危害关系，指病害和虫害会危害农作物的正常生长发育或者农药和肥料超出了适宜的范围，使得农作物的生长受到了影响。表现关系包括农作物对症状的表现关系，指农作物受到环境影响或者病虫害的危害时，会表现出相应的症状。促进关系包括农药、肥料、自然环境对农作物的促进关系，指合适的农药、肥料施在农作物上，促进了农作物的生长，或者在合适的自然环境下农作物能很好的生长。抑制关系包括自然环境对农作物的抑制关系，指恶劣的自然环境会对农作物的生长起到一定的抑制作用。导致关系包括病害、虫害、自然环境对症状的导致关系，指病害、虫害、自然环境对农作物产生影响后，农作物所表现出来的症状。表征关系包括症状对病害和虫害的表征关系，指根据农作物出现的症状，可以推断出影响农作物的病害或者虫害。诱发关系包括虫害对病害的诱发关系以及自然环境对病害和虫害的诱发关系，指病虫可以携带导致病害发生的病原，或者合适的自然环境下会促进病原或者病虫的生长。防治关系包括农药对病害和虫害的防治关系、虫害对虫害的防治关系以及肥料对病害、虫害、症状的防治关系，指喷洒合适的农药能够防止农作物受到病害或者虫害的危害，或者生物防治。缓解关系包括自然环境对症状、病害以及虫害的缓解关系，指在指定的环境下病害和虫害所导致的症状会有所缓解。恶化关系包括肥料对病害的恶化关系以及农药、肥料对虫害的恶化关系，指农药或者肥料使用不合适时会加重病害或者虫害对农作物的危害。上下位关系包括农作物与部位、型号、品种的上下位关系，指同一种农作物会有不同的型号或者品种，其都与最广泛的大类定义有上下位关系。

表3-2 农业语义关系

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 头实体  尾实体 | 农作物 | 病害 | 虫害 | 自然环境 | 农药 | 肥料 | 症状 |
| 农作物 | 上下位 | 抗性 | 抗性 | 抗性 |  |  | 表现 |
| 病害 | 危害 |  |  |  |  |  | 导致 |
| 虫害 | 危害 | 诱发 | 防治 |  | 抗性 |  | 导致 |
| 自然环境 | 促进、抑制 | 诱发、缓解 | 诱发、缓解 |  |  |  | 导致、缓解 |
| 农药 | 促进、危害 | 防治 | 防治  恶化 |  |  |  |  |
| 肥料 | 促进、危害 | 防治、恶化 | 防治、恶化 |  |  |  | 防治 |
| 症状 |  | 表征 | 表征 |  |  |  |  |

### 3.3.4 本体评估

本体评估标准一般包括清晰性、一致性、完善性和可扩展性[36]。清晰性指所定义的类和属性必须是明确的，不存在歧义；一致性指类间的关系在逻辑上必须是一致的；完整性指所定义的领域内类和属性是完整的，可以用于描述一个知识体系；可扩展性指当前领域出现新的概念时，本体可规模化的扩展[37]。本文本体初步构建完成后采用专家咨询的方法，邀请了农学的专家进行评估，最终设计的本体模型的概要图如图3-1所示。

## 3.4 本章小结

本章首先对用到的一些数据源进行介绍，包括《中国农作物病虫害》、《农业科学叙词表》等，然后提出了本文构建的本体设计，之后定义了本体设计里面的七类实体及其子实体和十二类语义关系。至此，本文完成了农业知识图谱模式层的本体模型构建，能够为之后数据层的搭建提供语义框架基础。

# 第4章 农业知识图谱数据层构建

## 4.1 引言

## 在完成了农业知识图谱模式层的构建后，本文基于农业知识图谱模式层构建的本体，对结构化和非结构化的数据进行挖掘，并将其存储在图数据库中，完成了农业知识图谱数据层构建工作。

## 本章将从以下几个方面介绍农业知识图谱数据层的构建工作：首先在第二小节介绍了农业知识图谱构建的整体流程；然后在第三小节介绍了构建农业实体数据集和农业关系数据集所用到的标注工具，以及标注完成后的格式；之后在第四小节介绍了挖掘非结构化数据知识所用到的实体识别模型，首先介绍了迁移学习领域里用到的一个利用自注意力机制和对抗学习方法的模型，然后介绍了本文所改进的基于知识蒸馏的迁移学习模型，并介绍了实验所用到的评价标准，数据参数以及实验结果等；随后在第五小节介绍了挖掘非结构化数据知识所用到的关系抽取模型，并且介绍了实验相关的一些信息；最后在第六小节介绍了挖掘完结构化和非结构化数据后，数据所存储的方式以及数据量等信息。

## 4.2 农业知识图谱构建流程

农业知识图谱构建包括模式层构建和数据层构建。其中数据层是由实体及实体间的关系构成的。本文采用的知识图谱的构建流程如图4-1所示。首先需要搜集构建图谱所需要的数据源，包括结构化和非结构化的数据，其中结构化的数据是高度组织和整齐格式的数据，可以来源于《农业科学叙词表》，而非结构化的数据本质上是结构化数据之外的一切数据，可以选择《中国农作物病虫害（第三部）》这类书本数据。然后需要从搜集到的数据中根据前面构建的本体模型来提取出相关的实体和关系，并将其表示成（实体，关系，实体）这样三元组的形式。对于结构化的数据，可以通过规则映射的方式，将其转换为我们所需要的三元组形式，对于半结构化和非结构化数据，可以采用深度学习的方法，从中抽取出本体模型中定义的实体以及实体间的关系。最后需要将前面抽取出来的三元组进行融合，并存储在neo4j图数据库中，以便后续查询和使用。

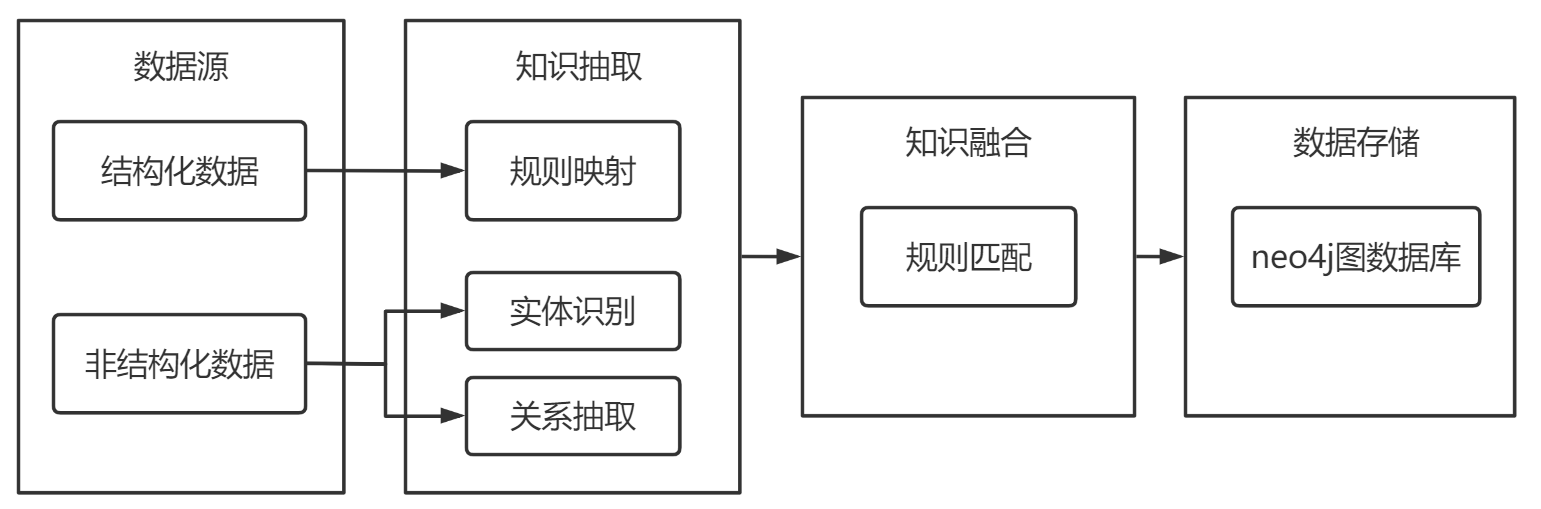


图4-1 农业知识图谱数据层构建流程

## 

## 4.3 农业数据标注工具

数据层构建过程中，对非结构化的数据需要用到深度学习的方法来进行知识抽取，而现存带有标注的农业数据很少，因此需要对非结构化的数据进行标注工作。为此，我们通过java编写了一个可视化的标注工具，以方便对文本数据按照本体模型的要求进行标注任务。该工具分为两部分，一部分是实体标注，另一部分是实体关系标注。



图4-2 实体标注界面

对于实体标注部分，其界面如图4-2所示。标注者可以通过上方的打开文件按钮，选择需要标注的文本。之后可以通过鼠标选出需要标注的实体，点击鼠标右键选择添加实体或者通过上方的添加实体按钮进行标注，标注的实体会出现在界面下方，这时可以点击实体后面的类型那一栏会出现一个下拉框，可以选择实体的类型，属性那一栏的下拉框则会根据已选择的实体类型展示相应的属性。对于标注好的实体，在上方文本中还将根据实体类型改变字体颜色。实体标注完成后可以点击导出结果按钮将标注结果进行导出到文件中，该工具还设有自动保存功能，在标注完实体后会将结果自动保存在文件中，以防意外情况未点导出结果按钮而关闭工具。导入NE按钮则是将之前标注的结果导入工具中，以方便继续进行标注任务。实体标注完后导出的文件将以ent作为文件后缀。以第一个标注的实体小麦为例，其在文件中表示为“C=小麦##P=14:16##T=大田作物##E=true”，其中C表示实体名称，P表示实体位置，T表示实体类型，若C中选择的内容为属性，则会用R来表示，最后E表示该段选择的内容是否为实体而不是实体的属性，所有内容用##进行分隔。



图4-3 实体关系标注界面

对于实体关系标注部分，其界面如图4-3所示。标注者可以通过上方的打开文件按钮选择之前标注过实体的文本，然后通过上方的导入NE按钮导入标注好的实体信息。之后通过上方的实体1和实体2按钮添加对应的头实体和尾实体，然后点击添加实体关系按钮即可在下方生产实体关系，由于头实体和尾实体类型确定后，实体关系也基本确定了，为了减小任务量，关系类型会默认给出对应的关系，当关系类型不是唯一的时候，则会给出相应的选项供标注者选择。实体关系标注完成后也可以点击导出结果按钮将标注结果进行导出到文件中，关系标注部分同样设有自动保存功能，以及导入关系功能。关系标注完后导出的文件将以rel作为文件后缀。以第一个标注的条锈病危害小麦为例，其在文件中表示为“C=条锈病##P=16:19##T=真菌病害||R=危害||C=小麦##P=14:16##T=大田作物”，其中R表示关系，实体和关系用||进行分隔，关于实体的表示方式与上述实体输出文档相同。

## 4.4 命名实体识别

农业命名实体识别任务是从非结构化数据中提取出农业知识图谱模式层中定义的实体，如大田作物、农药、肥料等，能构成知识图谱中的节点，是构建农业知识图谱中重要的一环。

由于搜集不到农业领域中符合本体模型中定义的带标签的数据，于是需要我们自己对搜集到的文本数据进行标注，考虑到标注数据的所需的人力和物力成本，本文打算采取迁移学习的思想来降低标注成本，同时利用知识蒸馏方法来提高目标域模型的效果。

4.4.1 自注意力机制下的对抗迁移学习模型

本文采用自注意力机制下的对抗迁移学习模型其结构如图4-4所示，总共分为五个部分：

（1）词嵌入层。该层将中文句子x = {c1, c2, ..., cN}中的每个字符ci都通过一个预训练好的嵌入矩阵将其映射嵌入向量。该部分输出为x = {x1, x2, ..., xN}。

（2）共享特征提取器和私有特征提取器。该部分由三个BiLSTM构成，分别用作源域私有特征提取器，源域和目标域共享特征提取器以及目标域私有特征提取器。以源域的数据输入为例，其过源域私有特征提取器后输出结果为，其中，表示源域私有特征提取器的参数；其过共享特征提取器后输出结果为，其中，表示共享特征提取器的参数。目标域的数据过特征提取器的结果同理。

（3）自注意力机制Self-Attention。该部分以特征提取器输出的结果作为输入，与Self-Attention中的三个参数相乘得到Q，K，V三个矩阵，于是得到，其中Q，该部分接下来采用Multi-Head Attention，即由多个Self-Attention结果拼接而成，最后该部分输出为，其中。

（4）CRF层。该层的输入由私有域中attention层出来的结果和共享域中attention层出来的结果拼接在一起，然后通过线性层得到句子中每个字对应标签的概率分布o = {o1, o2, ..., oN}，接着结合CRF层中的状态转移矩阵T给出句子的得分，其中表示第i个字取标签yi的得分，为标签yi-1到标签yi的转移得分。于是该部分的损失采用负对数似然，最终定义为，其中，其中Y表示句子所有可能取到的标签分布。

（5）领域鉴别器。该部分的输入是共享域中的attention层的输出结果，记为，该部分的输出结果为，其中表示鉴别器的参数，和是可学习的参数。于是该部分的对抗损失为，其中表示共享中的特征提取器的参数，表示共享域的特征提取器，K表示领域的数量，表示领域k中训练样本量。

最终，该模型的总损失定义为，其中是一个转换函数，当x属于源域时输出1，当x属于目标域时，输出0。

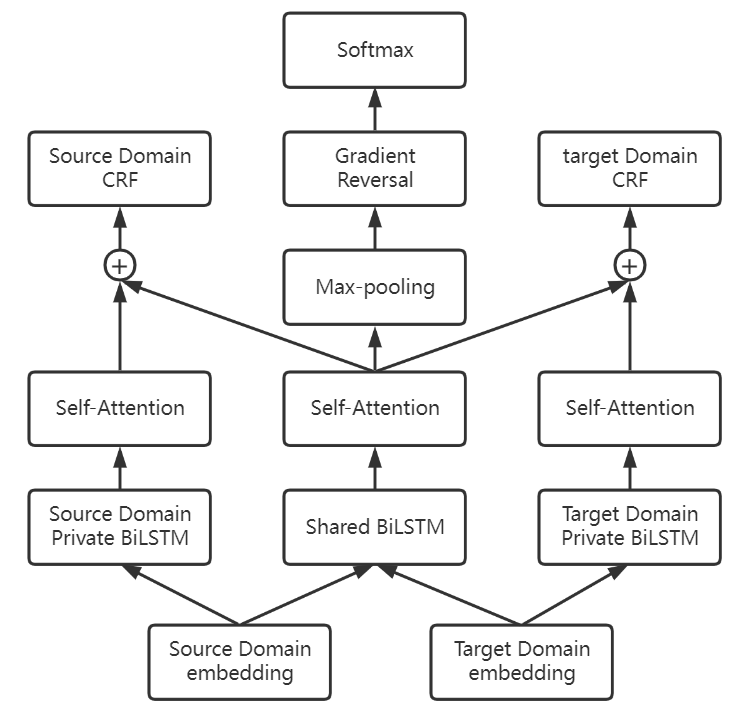


图4-4 自注意力机制下的对抗迁移学习模型

### 4.4.2 基于知识蒸馏的迁移学习模型

本文在BiLSTM-CRF的基础上，结合知识蒸馏的方法，提出了一种改进的实体识别模型，如图4-5所示。

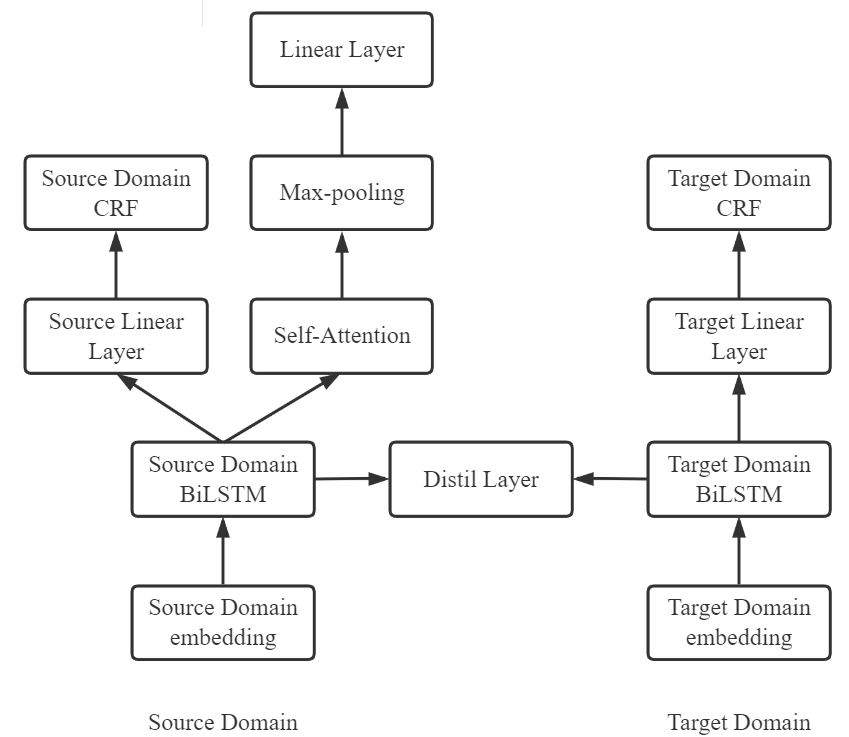


图4-5 基于知识蒸馏的迁移学习模型

模型训练过程分为如下几步：

（1）训练源域模型：将源域的中文句子中的每个字符都通过一个预训练好的嵌入矩阵将其映射嵌入向量。则源域的嵌入模块输出为。随后将输入到源域的特征提取器BiLSTM中，通过源域BiLSTM后输出结果为，其中，表示源域私有特征提取器的参数，de表示特征维度。然后通过线性层得到句子中每个字对应标签的概率分布o = {o1, o2, ..., oN}，接着结合源域CRF层中的状态转移矩阵T给出句子的得分，其中表示第i个字取标签yi的得分，为标签yi-1到标签yi的转移得分。于是该部分的损失采用负对数似然，最终定义为，其中，其中Y表示句子所有可能取到的标签分布。

（2）训练领域鉴别器：该部分以源域和目标域的数据通过源域的特征提取器BiLSTM后的结果作为输入，与Self-Attention中的三个参数相乘得到Q，K，V三个矩阵，于是得到，其中Q，该部分接下来采用Multi-Head Attention，即由多个Self-Attention结果拼接而成，最后自注意力机制部分的输出为，其中。于是领域鉴别器的判断结果为：d = ，其中表示鉴别器的参数，和是可学习的参数。领域鉴别器的损失采用交叉熵函数，其中是一个转换函数，当d属于源域时输出1，当d属于目标域时，输出0。

（3）数据筛选：当源域模型和领域鉴别器都训练完成后，将目标域中没有标签的文本数据作为输入，通过领域鉴别器将目标域数据分为S和T两部分，如图4-6所示，其中S部分为领域鉴别器识别为源域的数据，T部分为领域鉴别器识别为目标域的数据。

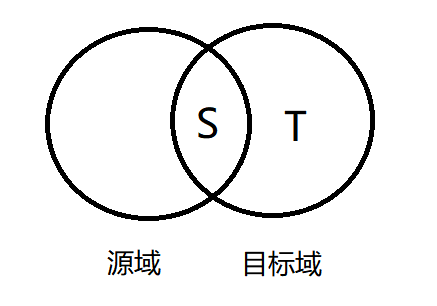


图4-6 目标域数据分割示意图

（4）知识蒸馏操作：将目标域的中文句子输入源域的嵌入模型以及源域的特征提取器后输出源域的特征，输入目标域的嵌入模型以及目标域的特征提取器后输出目标域的特征，随后将两个域特征输入知识蒸馏层，该网络层将目标域的特征通过多层感知机后得到目标域特征的均值和方差。最后知识蒸馏的损失定义为。目标域中有标签的数据的损失定义与训练源域模型的损失类似，。最后模型总的损失为。

### 4.4.3 实验结果及分析

实验结果中精确率（precision）、召回率（recall）以及调和平均值（F1值）的计算方式如下所示。

（4-1）

（4-2）

（4-3）

设百度百科的数据为源域，书本的数据为目标域。实验中的参数设置为训练数据的batch\_size为10，训练轮次eopch为500，优化器的优化方法设为Adam，学习率lr为0.01，权重衰减weight\_decay为0.0001。源域模型和目标域模型都是用BiLSTM-CRF模型。

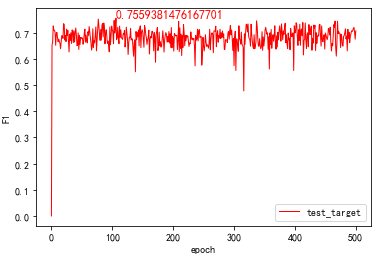
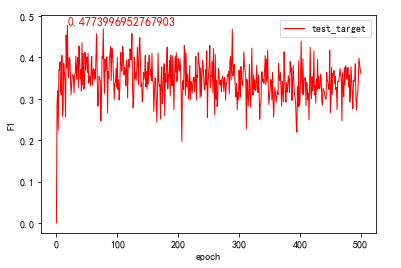
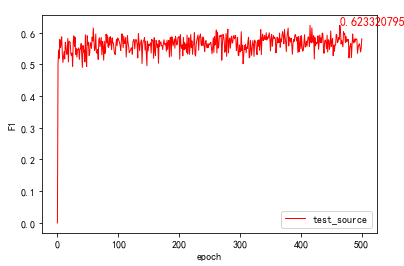


图4-7 目标域模型在目标域测试集上的F1值

目标域模型在用目标域所有训练集训练，在目标域测试集上测试，该模型的F1值随训练轮次变化如图4-7所示，可以看到目标域模型在训练300轮左右取到了最大值，为75.59%。



(b) 目标域测试集

(a) 源域测试集

图4-8 源域模型在不同域的测试集上的F1值

源域模型在用源域所有训练集训练，在源域测试集上测试，该模型的F1值随训练的轮次的变化如图4-8 (a)所示，可以看出，该模型在源域测试集上的最好效果为62.33%。源域模型在用源域所有训练集训练，在目标域上测试，该模型的F1值随着训练轮次的变化如图4-8（b）所示，可以看出，该模型在目标域测试集上最好的效果为47.74%。当该模型在源域测试集上取得最好效果时，在目标域测试集上的F1值为34.17%。

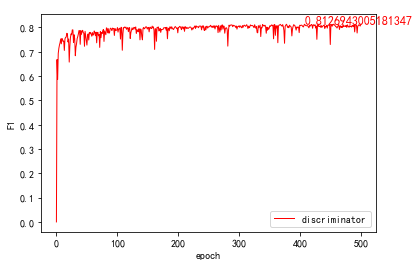


图4-9 领域鉴别器训练500轮的F1值变化图

领域鉴别器模型用源域和目标的数据混合训练，在源域和目标域混合数据进行测试，该模型的F1值随着训练轮次的变化如图4-9所示，可以看出该领域鉴别器最好的效果基本稳定在81.27%左右，仍有部分数据该领域鉴别器无法正确区分出其所属的领域。通过该领域鉴别器对目标域训练集数据（共计7303个）进行筛选，结果该领域鉴别器判断为源域的数据有425个，判断为目标域的数据有6878个。

对目标域训练集进行划分后，记目标域训练数据通过领域鉴别器后被识别为源域的部分记为S，被识别为目标域的部分记为T，则目标域模型在用不同的训练集训练以及不同的数据集进行知识蒸馏的实验结果如表4-1所示。其中测试集均为目标域测试集，目标域模型均为BiLSTM-CRF。该表中的每组实验结果取该组所有训练轮次中F1的最高值。

表4-1 不同训练数据和知识蒸馏数据的训练结果

|  |  |  |
| --- | --- | --- |
| 训练数据（带标签） | 知识蒸馏数据（不带标签） | F1值 |
| S+T | None | 75.83% |
| S+T | S | 76.02% |
| S+T | S+T | 73.88% |
| S+T | T | 72.93% |
| S | None | 52.65% |
| S | S | 64.57% |

从表4-1可以看出目标域模型在用所有目标域数据进行训练后的F1值是75.83%，加入S部分的数据进行知识蒸馏的时候，F1值为76.02%，效果提高了0.19%，但再加入T部分的数据进行知识蒸馏后，F1值为73.88%，效果反而下降了，说明S部分的数据对源域模型来说是有效的数据，能够提高目标域的模型，但T部分的数据通过源域模型会产生错误的特征，从而会影响目标域模型的效果。

与此同时，如果我们只用S部分的数据训练目标域，其F1值为52.58%，但如果加入S部分的数据进行知识蒸馏的时候，目标域的F1值提升到了64.16%，虽然说比用S+T的全部数据集训练的效果差些，但是S部分的数据只占了总体目标域数据的5.8%，极大的缩减了标注的任务量，从而极大的减少了标注的成本。

表4-2 loss中不同alpha对实验结果的影响

|  |  |  |  |
| --- | --- | --- | --- |
| alpha | precision | recall | F1值 |
| 0 | 0.698050 | 0.422689 | 0.526542 |
| 0.01 | 0.689452 | 0.438526 | 0.536079 |
| 0.02 | 0.669565 | 0.443421 | 0.533518 |
| 0.05 | 0.689333 | 0.446588 | 0.542023 |
| 0.1 | 0.665042 | 0.432191 | 0.523909 |
| 0.2 | 0.654726 | 0.436798 | 0.524007 |
| 0.5 | 0.682136 | 0.478261 | 0.562288 |
| 1 | 0.665140 | 0.627411 | 0.645725 |
| 2 | 0.743805 | 0.561762 | 0.640092 |
| 5 | 0.709355 | 0.591707 | 0.645212 |
| 10 | 0.722364 | 0.570112 | 0.637271 |
| 20 | 0.667824 | 0.498416 | 0.570816 |

目标域模型只用S部分的数据进行训练，用目标域测试集进行测试，该模型的精确率、召回率以及F1值随loss中alpha的变化如表4-2所示，可以看出，随着alpha的不断增加，目标域模型效果呈现出先增加后减少的趋势，其最好的效果在alpha取1至5区间内。当alpha变得越大时，知识蒸馏的影响就越大，当alpha的值超过一个阈值后，会使源域对目标域的影响过大，从而使得目标域模型的效果逐渐变差。而alpha越小时，知识蒸馏的影响就越小，则会使目标域模型的F1值趋于52.65%。

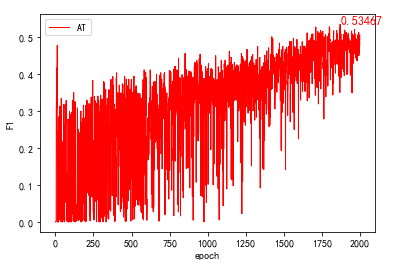


图4-10 领域鉴别器训练2000轮的F1值变化图

自注意力机制下的对抗迁移学习模型的数据集同样取百度百科的数据作为源域，目标域数据集取S部分进行训练，然后用所有目标域测试集进行测试，实验中的参数设置为训练数据的batch\_size为32，训练轮次epoch为2000，优化器的优化方法设为Adam，学习率lr为0.01，损失函数中alpha=0.06。最后模型的F1值随着训练轮次epoch的变化如图4-10所示，可以看出，该模型F1震荡比较激烈，且效果最好时为53.47%，相比于只用S部分进行训练的模型效果，该方法最终效果有所提升，但是提升不大。

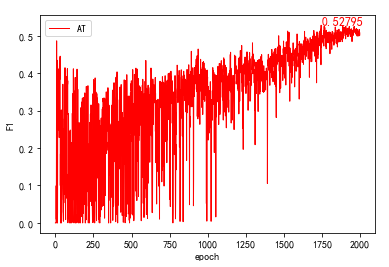


图4-11 领域鉴别器训练2000轮的F1值变化图

自注意力机制下的对抗迁移学习模型的数据集取公开数据集SIGHAN2005中的MSR数据集作为源域，目标域数据集取S部分进行训练，然后用所有目标域测试集进行测试，模型的F1值随着训练轮次epoch的变化如图4-11所示，该模型效果最好时为52.80%，可以看出在该模型中，非农业领域的公开数据集对农业领域的小部分数据集提升效果不大。

## 4.5 关系抽取

### 4.5.1 PCNN模型

本文知识蒸馏采用PCNN网络，其模型结构如图4-12所示。

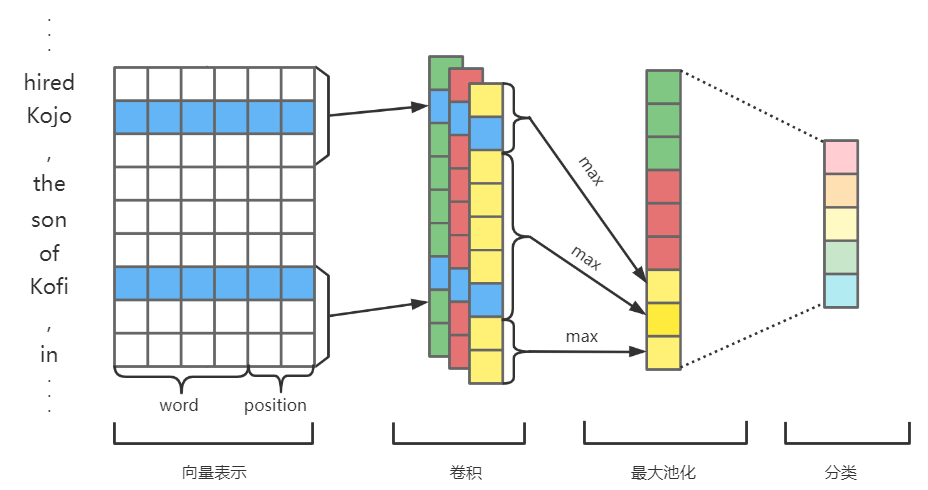


图4-12 PCNN模型

PCNN模型训练过程分为如下几步：

（1）数据预处理：首先对句子中的字进行位置编码，按句子中各个字离句子中实体的距离进行编码。由于句子中有两个实体，所以这条句子就会产生两个和句子长度相同的位置编码。于是句子中每个字的向量表示就是该字的嵌入表示和该字的两个位置编码进行拼接。

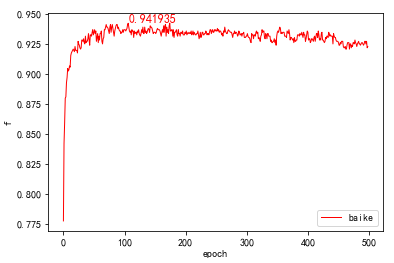
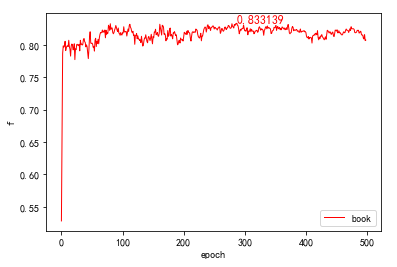
（2）特征提取：将数据预处理后的句子输入CNN 网络提取特征。

（3）关系分类：CNN网络层提取出来的特征根据两个实体的位置分为三段进行最大池化操作，拼接后输入线性层然后进行softmax操作得到句子中指定实体对的关系的分类结果。

### 4.5.2 实验结果及分析

实验结果中精确率（precision）、召回率（recall）以及调和平均值（F1值）的计算方式如公式所示，本实验计算的是微平均，因此根据计算公式可以得到精确率、召回率和F1值三者在数值上相同。

训练该模型参数设置为训练数据的batch\_size为32，训练轮次eopch为500，优化器的优化方法设为Adam，学习率lr为0.005，学习率衰减lr\_decay为0.95，权重衰减weight\_decay为0.0。



(b) 百度百科数据

(a) 书本数据

图4-13 PCNN在数据集上训练500轮的实验结果

用书本中标注的数据集进行训练时，该模型的f1值随训练轮次变化如图4-13(a)所示，可以看到，在训练100轮之后，f1值稳定在0.8至0.84之间，其中最高值为0.8331。用百度百科中标注的数据集进行训练时，该模型的f1值随训练轮次变化如图4-13(b)所示，可以看出，在训练完100轮左右，该模型的f1值达到最大，为0.9420，在之后的训练轮次中，基本稳定在0.925值0.942之间。

## 4.6 知识图谱存储方法

## 本文将知识抽取中抽取出来的三元组存储在Neo4j数据库中。Neo4j是一个高性能的NoSQL图形数据库，与一般的数据库不同，它不是将数据存储在表上，而是存在“图”数据结构上，其也可以被看作是一个高性能的图计算引擎，该引擎具有成熟数据库的所有特性。

## 知识图谱中所含各个类型的实体数量统计如表4-3所示，共计13983个实体，知识图谱中实体间所含各个类型的关系数量统计如表4-4所示，共计26498个关系，使用查询语句看到的以叶枯病为中心的一层关系子图如图4-14所示。

表4-3 知识图谱中实体数量统计

|  |  |
| --- | --- |
| 实体类型 | 数量 |
| 农作物 | 6913 |
| 病害 | 497 |

表4-3（续表）

|  |  |
| --- | --- |
| 实体类型 | 数量 |
| 虫害 | 779 |
| 农药 | 1511 |
| 肥料 | 473 |
| 自然环境 | 1980 |
| 症状 | 1830 |

表4-4 知识图谱中关系数量统计

|  |  |
| --- | --- |
| 关系类型 | 数量 |
| 抗性 | 2227 |
| 危害 | 4788 |
| 表现 | 3497 |
| 促进 | 2188 |
| 抑制 | 384 |
| 导致 | 2652 |
| 表征 | 2227 |
| 诱发 | 2328 |
| 防治 | 2479 |
| 缓解 | 855 |
| 恶化 | 100 |
| 上下位 | 2773 |

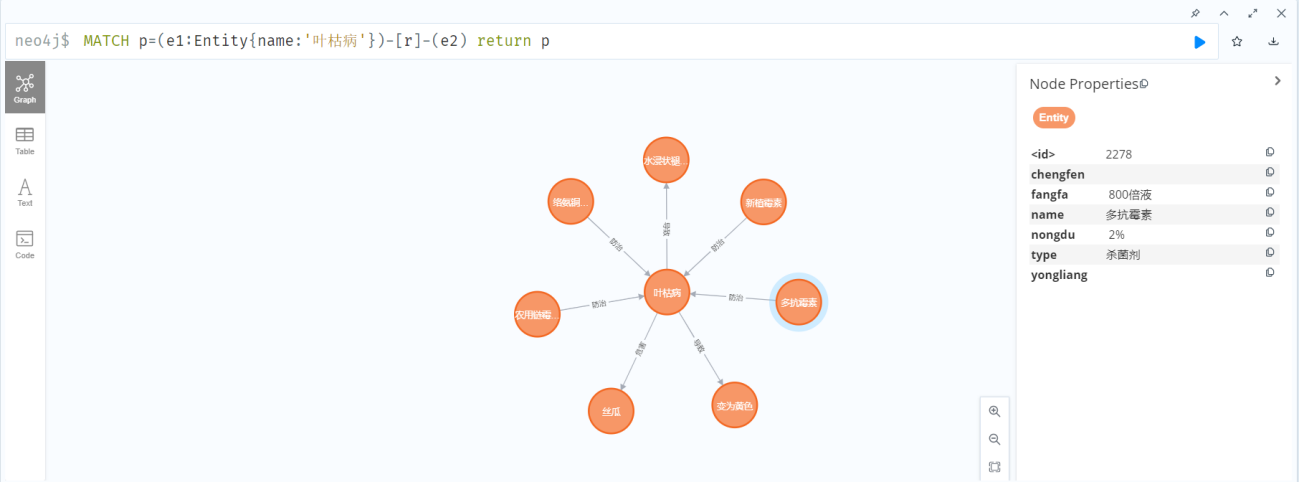


图4-14 以叶枯病为中心的一层关系子图

## 4.7 本章小结

本章在农业知识图谱模式层的基础上构建了知识图谱的数据层。在构建知识图谱数据层的过程中采用了深度学习的方法对非结构化的数据进行了挖掘，包括命名实体识别和关系抽取，由于训练深度学习模型需要大量的标注数据，而进行数据标注需要耗费大量的人力物力，因此，本章采用了基于知识蒸馏的迁移学习模型进行命名实体识别任务，实验结果发现，该模型在只有少量的目标域标注数据的情况下，能很好的提升目标域模型的效果，并且这少量的标注数据只占不到全体目标域数据的10%，极大的节省的人物和物力。随后本章采用PCNN模型进行关系抽取任务，发现其在测试集上的效果还不错。最后本章采用neo4j的图形数据库在存储所有的实体关系数据。至此，农业知识图谱的数据层构建完成。

# 第5章 农业知识图谱系统的设计与实现

## 5.1 引言

## 在完成农业知识图谱模式层和农业知识图谱数据层的构建工作后，依据模式层构建的本体以及数据层提供的三元组数据，完成了农业知识图谱构建工作，本文在其农业知识图谱的基础上设计并实现了农业知识图谱系统，其能让实体识别和关系抽取等技术更加直观，也能让我们通过查询功能或者问答功能来快速获得知识图谱中的信息。

## 本章将从以下几个方面介绍农业知识图谱系统的设计与实现工作：首先在第二小节介绍了系统的整体逻辑架构；然后在第三小节中介绍了知识图谱推理模型，并介绍了实验相关的一些信息；最后在第四小节中介绍了系统中各个功能模块的实现，之前数据层构建过程中用到的实体识别等技术也将其应用到农业知识图谱平台中。

## 5.2 系统逻辑架构

本系统集成了构建农业知识图谱所需的一些技术，用于展示农业知识图谱构建及应用等模块，包括实体识别，关系抽取，知识推理等。为了应对未来规模的不断扩大，本系统采用图5-1所示的系统逻辑结构，能够随时根据需求进行扩展。

本系统一共分为四层，从下往上依次是数据层，模型层，应用层和展示层。其中数据层是农业知识图谱数据产生的过程，目前构建知识图谱的数据源主要来自书本数据和网上的百科数据，这些非结构化数据通过实体识别和关系抽取等机器学习模型提取出构建知识图谱所需的数据，然后存储在neo4j图形数据库中。模型层则包括知识图谱构建以及扩展过程中所需要用到的各类模型，目前包括实体识别模型，关系抽取模型，知识推理模型等等。应用层包括各类模型的应用以及知识图谱的相关应用等，随着后续数据和技术的不断扩充，还可能包含自动回答农民问题的自动问答系统等。展示层则指的是系统本身的载体，本文使用web网站作为载体，将来源不同的应用（比如python编写的实体识别模型，以及java编写的知识图谱查询系统）集中在一起，用户可以根据自身需求跳转至对应的应用界面。

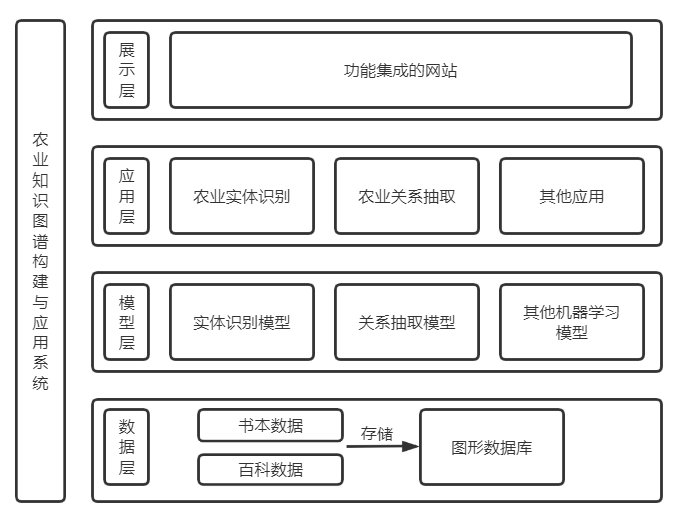


图5-1 系统逻辑结构图

## 

## 5.3 知识图谱推理

### 5.3.1 TuckER模型

在知识图谱表示学习的模型中，我们使用了一种基于Tucker分解的模型（Tucker分解的公式如式（5-1）所示）。

式（5-1）

在该模型中，取A和C为实体嵌入矩阵E，B为关系嵌入矩阵R，即，，其中ne和nr分别表示实体和关系的数量，de和dr分别表示实体和关系的嵌入维度。于是，该模型定义的得分函数如式（5-2）所示，其中是头实体和尾实体的嵌入表示，是实体嵌入矩阵E的某一行，是是关系的嵌入表示，即关系嵌入矩阵R的某一行，是张量。

式（5-2）

在随后的训练过程中，将训练集的三元组通过得分函数计算出得分后，通过sigmoid函数得到一个概率值，表示该三元组是正确的一个概率，最后通过logistic回归进行分类。

由于参数的数量只与实体和关系的嵌入维度相关，与实体和关系的数量无关，因此该模型的参数量与实体和关系的嵌入维度和线性相关。因为模型中有核张量，因此不像DistMult和SimplE那样，该模型没有把所有学到的知识都融入embedding中，而是把其中一些是融在核张量中，在所有的实体和关系中共享。

### 5.3.2 实验结果及分析

衡量知识图谱嵌入(Knowledge Graph Embedding)模型性能常用的指标有MRR，MR，HITS@1，HIT@3，HIT@10。其中MRR和HIT@10是两个重要指标。

对于测试集中每一个三元组(h , r , t)，此处以预测尾实体t为例，我们用知识图谱中的每一个实体ti来替换这个尾实体t，然后将新的三元组送入得分函数算出该三元组的得分，这样我们就能得到一系列的分数，将这些分数从高到低进行排序，尾实体被替换前的三元组的得分在这一系列得分中的排名记为rank，该排名越靠前，即rank值越小，模型效果越好。例如，对于测试集中的三元组（条锈病，危害，小麦），用知识图谱中的所有实体替换掉尾实体小麦后组成新三元组，将新三元组送入得分函数算出各自的得分，并进行排序后，排名前5的三元组如表5-1所示，可以看到三元组（条锈病，危害，小麦）排在第二位，故测试集中三元组（条锈病，危害，小麦）的rank值为2。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 头实体 | 关系 | 尾实体 | 得分 | rank |
| 条锈病 | 危害 | 水稻 | 0.789 | 1 |
| 条锈病 | 危害 | 小麦 | 0.742 | 2 |
| 条锈病 | 危害 | 玉米 | 0.641 | 3 |
| 条锈病 | 危害 | 大豆 | 0.426 | 4 |
| 条锈病 | 危害 | 冬瓜 | 0.254 | 5 |
| ... | ... | ... | ... |  |

表5-1 测试用例三元组（条锈病，危害，小麦）替换尾实体得分情况

MRR的全称是Mean Reciprocal Ranking，具体计算方式如式（5-1）所示，其中S是测试集中三元组的集合，|S|是三元组集合个数，ranki是指第i个三元组的链接预测排名。通过公式可以看出，rank越小，MRR值越大，模型就效果越好。

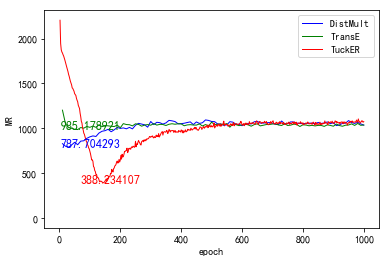
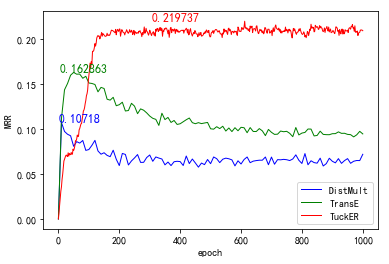
（5-1）

MR的全称是Mean Rank，具体计算方式如式（5-2）所示，其中涉及的符号和MRR计算公式中涉及的符号一样。通过公式可以看出，rank越小，MR值越小，模型效果就越好。

（5-2）

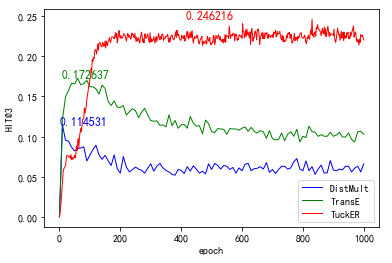
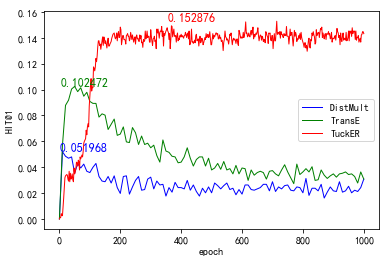
HITS@n指的是在链接预测中排名小于等于n的三元组平均占比，具体计算方式如式（5-3）所示，其中I(x)函数是indicator函数（当x为真时其输出为1，当x为假时其输出为0），其余涉及的符号和MRR计算公式中涉及的符号一样。一般地，取n等于1、3或者10。通过公式可以看出，rank小于等于n的三元组数量越多，HIT@n的值就越大，模型的效果就越好。

（5-3）



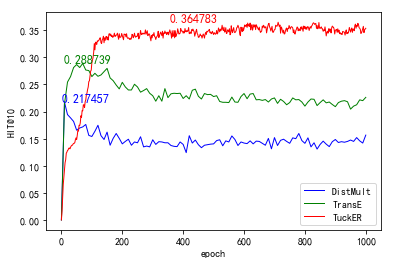
(b) MR

(a) MRR



(d) HIT@3

(c) HIT@1



(e) HIT@10

图5-2 DistMult、TranE以及TuckER模型训练1000轮的实验结果

该模型训练参数设置为训练数据的batch\_size为128，训练轮次eopch为500，优化器的优化方法设为Adam，学习率lr为0.005，实体向量的嵌入维度为200，关系向量的嵌入维度为200。

如图5-2所示是DistMult、TranE以及TuckER模型在农业数据集上的实验结果。这三个模型的MRR值随训练轮次变化如图5-2（a）所示，可以看到TuckER模型的MRR值最大，为0.2197，其次是TranE模型，其MRR值最大为0.1629，最后是DistMult模型，其MRR值最大为0.1072。这三个模型的MR值随轮次变化如图5-2（b）所示，TuckER模型的MR值最小为388.234。HIT@1随轮次变化如图5-2（c）所示，TuckER模型的HIT@1的最大值为0.1529。HIT@3随轮次变化如图5-2（d）所示，TuckER模型的HIT@3的最大值为0.2462。HIT@10随轮次变化如图5-2（e）所示，TuckER模型的HIT@10的最大值为0.3648，由于知识图谱中指定的头实体和关系可能存在多个尾实体，于是模型预测的尾实体得分可能并不会排在前3，但也是正确的三元组，因此一般用HIT@10来表示模型预测的效果。除此之外，TuckER模型在验证集（训练集中取的部分数据）上HIT@1，HIT@3以及HIT@10都为1.0，此处可以说明该模型能作为知识图谱的知识表示。

## 5.4 系统主要功能模块的实现

### 5.4.1 实体识别及关系抽取功能

该功能模块是基于python进行开发的，主要应用包括分词，词性标注，命名实体识别，关系抽取这几类。用户可以在上方待识别的句子中输入需要进行识别的文本，然后点击提交按钮，该模块则会将输入的文本稍加处理后送给深度学习模型，模型根据输入的文本计算出识别结果，然后将识别结果传回给应用平台，随后平台向用户展示出分词结果，词性标注结果，该文本中所包含的命名实体以及这些实体间存在的关系，该过程如图5-3所示。

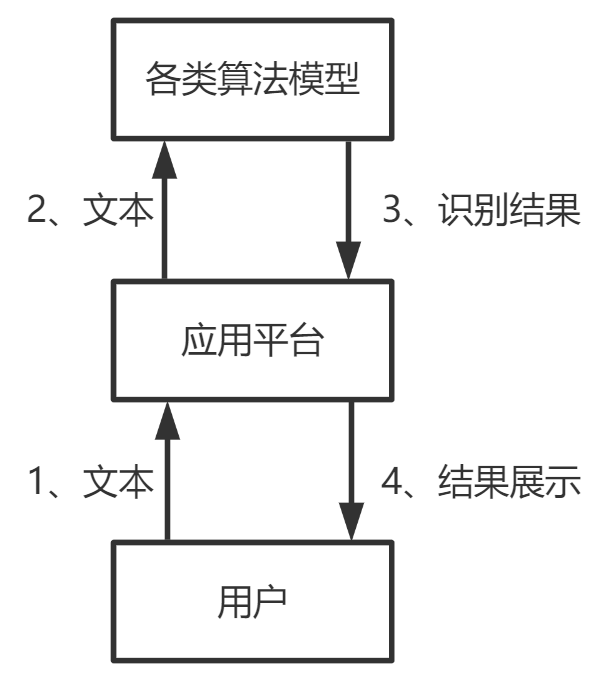


图5-3 系统调用模型过程



图5-4 分词，词性标注以及命名实体识别结果界面

结果界面展示如图5-4和图5-5所示。分词结果用符号“[]”进行分隔。词性标注结果中，是在分词结果的基础上给出每个词所对应的词性。命名实体识别结果中，逗号之前的表示命名实体，逗号之后的表示该实体对应的实体类别，根据不同的实体类别展示不同的颜色。关系抽取则被逗号分为三部分，其中从左往右第一部分表示头实体，第二部分表示尾实体，第三部分表示这两实体间的关系，然后根据不同的关系类型展示不同的颜色。可以从图5-4中看出，对于句子“炭疽病危害症状：炭疽病在果实成熟时危害果实，腐烂散发酒味。防治方法：清除落果病果，幼果期用70%甲基托布津800倍液或用世高喷施。”，实体识别模型识别出了炭疽病，甲基托布津以及世高等实体，并标上了对应的实体类别，从图5-5中可以看出，关系抽取模型也识别出了实体甲基托布津与实体炭疽病之间的防治关系。

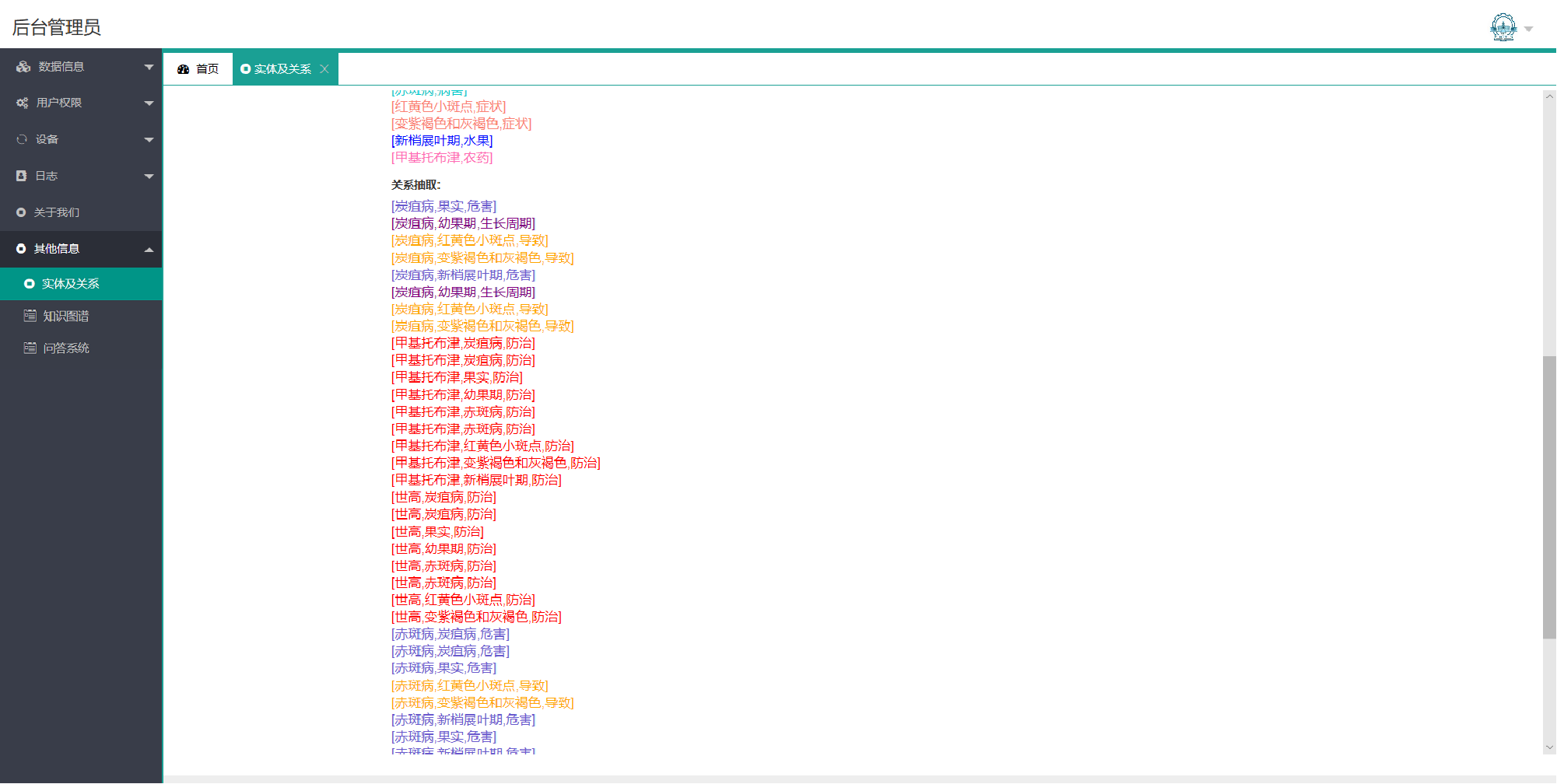


图5-5 分词，词性标注以及命名实体识别结果界面

### 5.4.2 知识图谱查询功能

该功能模块是基于java进行开发的，主要功能包括知识图谱的展示以及查询功能。该查询能够支持指定实体名称或者指定实体类别或者指定关系类型的一层关系和二层关系（包含正向关系和反向关系）查询，该功能模块还能够支持精确查询和模糊查询，当用户不太清楚想查询的实体的具体名称时，可以通过模糊查询来返回相近的结果。如图5-6所示，其为敌百虫粉剂的一个二层关系示意图，可以看出敌百虫粉剂能防治潜叶蝇，而潜叶蝇会危害西瓜等作物，因此图谱中这些关系表示可以用敌百虫粉剂来防治危害西瓜的潜叶蝇。

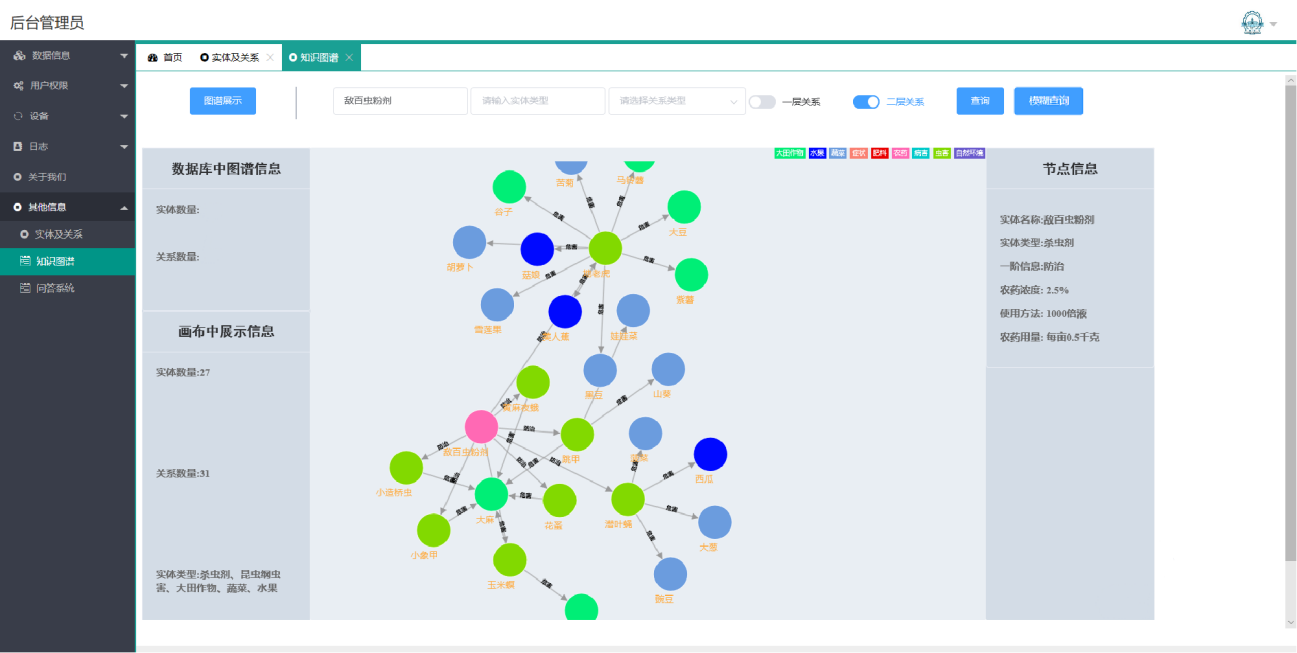


图5-6 知识图谱查询结果界面

### 

### 5.4.3 知识图谱问答功能

该功能模块主要是实现问答功能。用户可以在上方请输入问题中输入需要询问的问题，然后点击提交按钮，该模块则先会读取图数据库，将所有的实体和关系生产一个列表，将其加入jieba的词库中，然后把输入的问题进行通过jieba进行分词和词性标注，找出其中的实体和关系，随后运用基于规则的问题分类以及问答生成方法，即根据农业领域常见的问题设置规则模板，对问题进行分类以及问句模板的匹配，判断问题处理后的结果是否于之前设定的一些规则模板相匹配，若匹配成功则根据匹配的规则转换成图数据neo4j的查询语句，然后用查询语句对neo4j数据库进行查询，将返回的结果格式稍加处理后展示到前端。

如图5-7所示，当用户在请输入问题的框中输入问题“什么药能治叶锈病？”后，点击提交按钮，后端则会将问题进行分词和词性标注，识别出实体是叶锈病，关系是治，随后生成查询语句“MATCH (n)-[r:Relation{name:'防治'}]-(m:Entity{name:'叶锈病'}) return n”，将该查询语句输入neo4j数据库中进行查询得到返回结果，将结果稍作处理后传回前端答案的那个框中展示结果。

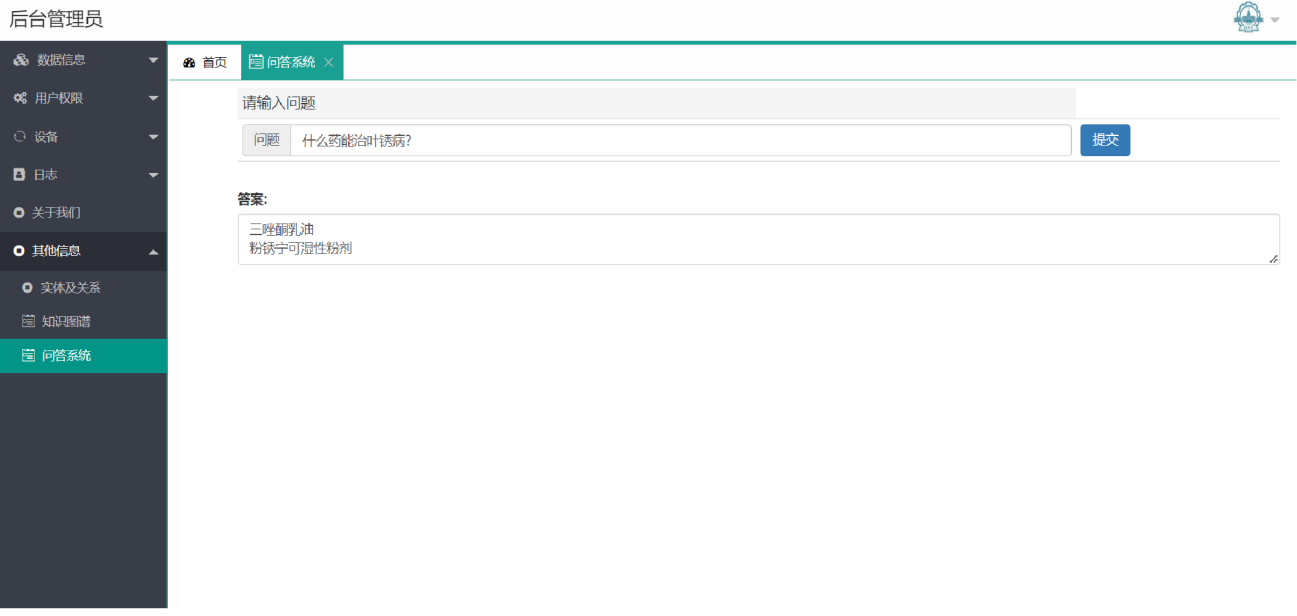


图5-7 知识问答功能回答结果界面

### 5.5 本章小结

本章训练了一个基于表示的知识推理模型，该模型可以用来表示农业知识图谱中现存的实体和关系，也可以基于农业知识图谱用来推理知识图谱中缺失的三元组是否是正确的，从而达到知识图谱补全的效果。接下来本章最终设计并实现了农业知识图谱系统，该系统包含各个功能模块，包括实体识别与关系抽取功能，知识图谱查询功能以及知识图谱问答功能。

# 结 论

本文对农业知识图谱构建技术进行了研究，首先基于农业领域现有的数据设计了本体以及语义加工规范，然后对农业领域的非结构化数据进行命名实体识别和关系抽取，搭建了知识图谱，在进行命名实体识别任务中，本文提出了基于知识蒸馏的迁移学习模型，该模型能够在目标域只有少量标注数据的情况下，通过源域模型的知识对目标域模型进行知识蒸馏，来提高目标域模型的效果。最后本文在搭建的农业知识图谱的基础上，设计并实现了农业知识图谱系统，该系统让我们快速获得知识图谱中的信息。

本论文的主要创造性工作归纳如下：

（1）提出了基于知识蒸馏的迁移学习模型，该模型能有效减少需要标注的数据量，达到减少人力物力的效果。该模型能够在目标域只有少量标注数据的情况下，通过源域模型的知识对目标域模型进行知识蒸馏，来提高目标域模型的效果。通过实验验证了该模型在农业数据集上通过知识蒸馏的方法能提高目标域只用少量数据训练的模型的效果。

（2）提出了农业领域的本体建模。该本体设计定义了知识图谱的抽象层表示，共计七类实体和十二类关系。

今后还应在以下几个方面继续深入研究：

（1）本文所用的农业知识图谱系统仅是采用了将深度学习模型存在服务器中，需要用时加载保存好的模型来进行推理预测任务，在以后的研究中，可以采用triton相关的技术来管理和应用农业知识图谱系统中的深度学习模型。

（2）本文所用的农业知识图谱问答模块仅是采用了基于规则的匹配方法，在以后的研究中，可以采用引入深度学习相关的技术，更好的分析处理用户提出的问题，以及更加符合上下文环境的自然语言回答。

（3）本文所用提出的模型的知识蒸馏方法可以尝试其他效果更好的知识蒸馏方法，以及领域鉴别器除了正常的识别数据所属的领域外，还可以尝试用对抗的方式。

# 参考文献

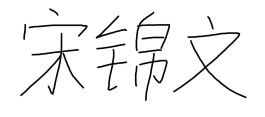
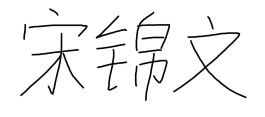
1. Krupka G, IsoQuest K. Description of the nerowl extractor system as used foe muc-7[C]. Proceedings of the 7th Message Understanding Conference, Virginia, 2005:21-28.
2. Black W J, Rinaldi F, Mowatt D. FACILE: Description of the NE System Used for MUC-7[C]. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
3. Appelt D E, Hobbs J R, Bear J, et al. SRI International FASTUS system: MUC-6 test results and analysis[C]. Proceedings of the 6th conference on Message understanding, 1995:237-248.
4. Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3):361-365.
5. Kapur J N. Maximum-entropy models in science and engineering[M]. New Jersey:John Wiley & Sons, 1989.
6. Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4):18-28.
7. Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1):81-106.
8. Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. Eighteenth International Conference on Machine Learning, 2001:282-289.
9. Lecun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4):541-551.
10. Elman J L. Finging Structure in Time[J]. Cognitive science, 1990, 14(2):179-211.
11. Hochreiter S, Schnidhuber J. Long Short-term Memory[J]. Neural computation, 1997, 9(8):1725-1780.
12. Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014:1724-1734.
13. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging [J]. CoRR abs/1808.01991, 2015.
14. Canasai Kruengkrai , [Thien Hai Nguyen](https://dblp.org/pid/130/6395.html), [Sharifah Aljunied Mahani](https://dblp.org/pid/84/9013.html), et al. Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling[C]. Proceedings of the 58rd Annual Meeting of the Association for Computational Linguistics, 2020:5898-5905.
15. [Kevin Clark](https://dblp.org/pid/78/6661.html), [Minh-Thang Luong](https://dblp.org/pid/12/8213.html), [Christopher D. Manning](https://dblp.org/pid/m/ChristopherDManning.html), et al. Semi-Supervised Sequence Modeling with Cross-View Training[C]. [Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018](https://dblp.org/db/conf/emnlp/emnlp2018.html" \l "ClarkLML18):1914-1925.
16. Marek Rei. Semi-supervised Multitask Learning for Sequence Labeling[C]. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017:2121-2130.
17. Zihan Liu, Genta Indra Winata, and Pascale Fung. Zero-Resource Cross-Domain Named Entity Recognition[C]. In Proceedings of the 5th Workshop on Representation Learning for NLP, 2020:1-6.
18. Bari, M. S., Joty, S., and Jwalapuram, P. Zero-Resource Cross-Lingual Named Entity Recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020:7415-7423.
19. [Pengfei Cao](https://dblp.org/pid/182/7941.html), [Yubo Chen](https://dblp.org/pid/90/7879.html), [Kang Liu](https://dblp.org/pid/42/4903.html), [Jun Zhao](https://dblp.org/pid/47/2026-1.html), et al. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism[C]. [Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018](https://dblp.org/db/conf/emnlp/emnlp2018.html" \l "ClarkLML18):182-192.
20. Aone, Chinatsu, Halverson, Lauren, Hampton, Tom, et al. SRA:Description of the IE2 System Used for MUC-7[C]. Proceedings of the 7th Conference on Message Understanding. 1998.
21. J. Fukumoto, M. Shimohata, F. Masui, and M. Saski. 1998. [Description of the Oki System as Used for MET-2](https://aclanthology.org/M98-1020). In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998.
22. Kambhatla, Nanda. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction[C]. Proceedings of the 42th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. 2004:178-181.
23. Miao, Qingliang, Zhang, Shu, Zhang, et al. Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text[C]. Proceedings of the 26th Pacific Asia Conference on Language , Information, and Computation. 2012:99-107.
24. Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. [Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks](https://aclanthology.org/D15-1203). In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1753–1762.
25. Pengda Qin, Weiran Xu, and William Yang Wang. [DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction](https://aclanthology.org/P18-1046). In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018:496–505.
26. Sun T, Zhang C, Ji Y,et al. Reinforcement learning for distantly supervised relation extraction[J]. IEEE Access 7, 2019:98023-98033.
27. Bordes A, Usunier N,Garcia-Duran A,et al. Translating Embeddings for Modeling Multi-Relational Data[C]. Neural Information Processing Systems (NIPS), 2013:1-9.
28. Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2014.
29. Lin Y, Liu Z, Sun M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2015.
30. Ji G, He S, Xu L, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[C]. Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, 2015:687-696.
31. Nickel M, Tresp V, Kriegel H P. Factorizing Yago: Scalable Machine Learning for Linked Data[C]. Proceedings of the 21st international conference on World Wide Web, 2012:271-280.
32. Yang B, Yih W T, He X, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases[J]. arXiv preprint arXiv:1412.6575, 2014.
33. [Ivana Balazevic](https://aclanthology.org/people/i/ivana-balazevic/), [Carl Allen](https://aclanthology.org/people/c/carl-allen/), [Timothy Hospedales](https://aclanthology.org/people/t/timothy-hospedales/). [TuckER: Tensor Factorization for Knowledge Graph Completion](https://aclanthology.org/D19-1522.pdf). [Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing](https://aclanthology.org/volumes/D19-1/), 2019:5185-5194.
34. Michael Schlichtkrull , Thomas N, Kipf, Peter Bloem, et al. Modeling Relational Data with Graph Convolutional Networks[C]. European semantic web conference, 2018:593-607.
35. Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel. Convolutional 2D Knowledge Graph Embeddings[C].
36. 李景，孟连生. 构建知识本体方法体系的比较研究[J]. 现代图书馆情报技术，2004：17-22.
37. 李悦. 农作物病虫害知识图谱构建研究[D]. 北京：中国农业科学院，2021.
38. 徐江南. 面向法律文本的命名实体识别研究[D]. 昆明：云南财经大学，2021.
39. 王鑫雷. 基于中文知识图谱的问答系统研究[D]. 大连：大连理工大学，2022.
40. 吴茜. 基于知识图谱的农业智能问答系统设计与实现[D]. 厦门：厦门大学，2020.
41. 白宇轩. 基于迁移学习的命名实体识别研究[D]. 南京：南京邮电大学，2022.
42. 乔波. 基于农业叙词表的知识图谱构建技术研究[D]. 长沙：湖南农业大学，2021.
43. 杨泽. 中国古典文学文本的命名实体识别及知识图谱构建研究[D]. 南京：南京邮电大学，2022.
44. 杨荣莹. 基于迁移学习的命名实体识别研究[D]. 贵阳：贵州大学，2022.
45. 赵旭. 基于迁移学习的中文命名实体识别研究[D]. 沈阳：吉林大学，2022.
46. 盛剑. 迁移学习在命名实体识别中的应用[D]. 哈尔滨：哈尔滨工业大学，2019.
47. 陈云鹏. 基于深度学习的关系抽取研究[D]. 沈阳：吉林大学，2021.
48. 徐娇. 基于注意力机制和图卷积网络的关系抽取应用研究[D]. 南京：南京邮电大学，2022.
49. 刘欢. 面向复杂文本场景的关系抽取方法研究[D]. 南京：东南大学，2021.
50. 张浩哲. 基于元学习的领域知识图谱补全技术[D]. 南京：南京邮电大学，2022.
51. 陈尚志. 基于事实信息融合的知识图谱补全方法研究与实现[D]. 济南：山东大学，2021.

# 攻读硕士学位期间取得创新性成果

# 哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于知识蒸馏的农业知识图谱构建研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。



作者签名： 日期：2022年6月22日

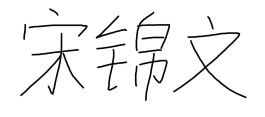
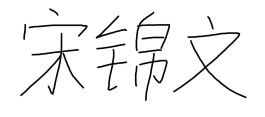
学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。



作者签名： 日期：2022年6月22日



导师签名： 日期：2022年6月22日

# 致 谢

衷心感谢导师王玲教授对本人的精心指导。她的言传身教将使我终身受益。

感谢姜京池老师提出思想和方法，以及实验室全体老师和同窗们的热情帮助和支持！

感谢东北农业大学的各位老师学生提出的指导意见以及参与的数据标注工作。

感谢实验室提供的经费以及训练模型用的服务器。

感谢魏博同学在我学习期间对我的帮助和支持。