

面向食品监管领域的知识图谱构建研究

卜意磊¹, 庞文迪¹, 吴甜甜², 杜奕坤³, 李珊^{2*}

(1. 江苏省工商行政管理局信息中心, 江苏 南京 210019;

2. 北京化工大学 信息学院, 北京 100029;

3. 南京航空航天大学 经济与管理学院, 江苏 南京 211100)

摘要: 该文提出了面向食品监管领域的知识图谱构建研究, 通过整理食品监管领域的相关文件和政策, 并进行实体识别、实体关系识别、实体对齐构建食品监管领域知识图谱。其中基于双向长短时记忆网络与条件随机场结合的 BiLSTM-CRF 模型进行实体识别, 准确率达 0.96; 基于食品监管实体的归类结果, 确定同标签的实体间的分类关系, 并创建“文本-实体”矩阵, 提取出包含某实体对的所有句子, 归纳实体对之间的关系; 通过聚类进行实体对齐, 并基于 Neo4j 存储和呈现图谱。构建的食品监管知识图谱弥补了食品监管领域知识图谱研究的空白, 提升了食品监管体系和监管能力现代化水平。

关键词: 知识图谱; 食品监管; 实体识别; 关系抽取; 实体对齐

中图分类号: TP391.1

文献标识码: A

文章编号: 1673-629X(2023)06-0202-06

doi:10.3969/j.issn.1673-629X.2023.06.030

Research on Knowledge Graph Construction for Food Supervision

BU Yi-lei¹, PANG Wen-di¹, WU Tian-tian², DU Yi-kun³, LI Shan^{2*}

(1. Jiangsu Provincial Administration for Industry and Commerce Information Center, Nanjing 210019, China;

2. School of Information, Beijing University of Chemical Technology, Beijing 100029, China;

3. School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China)

Abstract: Research on knowledge graph construction for the field of food supervision is proposed. By sorting out relevant documents and policies in the field of food supervision, and performing entity identification, entity relationship identification and entity alignment, a knowledge graph in the field of food supervision is constructed. Among them, the BiLSTM-CRF model based on the combination of bi-directional long short-term memory network and conditional random field is used for entity recognition, with an accuracy rate of 0.96. Based on the classification results of food regulatory entities, the classification relationship between entities with the same label is determined, and a "text-entity" matrix is constructed to extract all sentences containing a certain entity pair and summarize the relationship between entity pairs. Entity alignment is performed through clustering, and graphs based on Neo4j are stored and presented. The food supervision knowledge map constructed fills the gap of knowledge map research in the field of food supervision, and improves the modernization level of the food supervision system and supervision capacity.

Key words: knowledge graph; food supervision; entity recognition; relation extraction; entity alignment

0 引言

“十四五”时期是全面建成小康社会的关键时期, 是全面落实国家治理体系与治理能力现代化的推进期。对政府来说, 加强和改善市场监督管理体制是维护市场公平竞争、充分激发市场活力和创造力的重要保障, 是政府职能转变的重要方向。在此背景下, 市场监督管理局需要顺应时代变革的需求, 创新市场监管

工作, 从服务方法和形式上寻求突破。过去, 食品监管领域常使用表格的形式来管理、存储和展示监管数据。但进入网络与大数据时代后, 随着监管范围逐渐扩大、监管数据大幅增加, 表格显现出难以处理大量信息、难以及时发现新的知识等问题。尤其是电子商务与网络交易平台的主流化带来的跨地域交易等情况, 给食品监管带来很大的困难。为解决这一问题, 需要加强食

收稿日期: 2022-06-09

修回日期: 2022-10-10

基金项目: 国家自然科学基金面上项目(72174086)

作者简介: 卜意磊(1988-), 女, 硕士, 高工, 研究方向为数据分析、系统架构、政务信息化; 通讯作者: 李珊(1977-), 女, 博士, 副教授, 研究方向为数据分析。

品安全的信息化监管,推动监管形式的变革^[1]。

与传统的表格存储不同,知识图谱可以实现知识的规范存储,将数据保存为三元组 (Triple),并以图的形式展现出来。构建食品监管领域的知识图谱,可以直观地展示监管机构、食品企业等实体的属性信息及关系;通过知识图谱智能搜索的应用,可以帮助监管者便捷地检索和管理相关对象;运用知识图谱的知识推理等方法,可以辅助管理者发现监管中的潜在信息。该文以食品监管领域为研究对象,首先利用爬虫获取食品监管相关的规章制度、政策公文等文件,然后基于 BiLSTM-CRF 模型进行实体识别,接着通过归类和构建“文本-实体”矩阵的方式进行关系抽取,最后对食品监管实例进行去重和消歧,并使用 Neo4j 进行图谱存储和展示。提出食品监管知识图谱的实现流程,弥补了食品监管领域知识图谱研究的空白,为后续的语义搜索、智能问答、精准监管提供支撑,全面提升食品监管体系和监管能力现代化水平。

1 国内外研究现状

知识抽取包括命名实体识别及实体关系抽取。

命名实体识别 (Named Entity Recognition, NER) 的主要技术方法有三种,即基于规则和词典的方法、基于统计的方法,以及混合方法^[2]。随着深度学习技术的发展,神经网络模型也逐漸在构建知识图谱、进行知识抽取的任务中得以广泛应用。Huang 等^[3]提出了使用双向长短期记忆网络 (BiLSTM) 进行抽取的同时,增加 CRF 层来优化结果的新模型,即 BiLSTM-CRF。该模型较传统方法有了很大提升,也在其他研究中应用广泛。叶蕾等人把 BiLSTM-CRF 模型应用到中文电子病历的命名实体识别中,实现了长距离的依赖关系的获取^[4]。罗熹等人将此模型应用到中文临床实体识别,并结合了“多头自注意力机制”,捕捉字符的潜在依赖权重、语义关联等特征^[5]。

关系抽取在早期主要包括基于规则、字典、本体的方法;引入特征后出现了机器学习方法,包括监督、半监督、无监督等方式;为实现特征的自动提取,关系抽取开始使用深度学习方法。Hashimoto K 等首先提出用递归神经网络 RNN 进行实体关系抽取,该方法在语法树上运行,允许对目标重要短语显式加权,取得较好的结果^[6]。Zeng 等人提出了用更简单的卷积神经网络 CNN,无需复杂预处理即可实现实体关系抽取^[7]。Katiyar 等提出神经网络联合模型 BiLSTM+Attention,同时抽取实体、关系以及关系类型^[8]。孙劭芃等提出一种基于 BERT-Bi LSTM-Attention 的食品安全领域实体关系抽取模型,在测试集上实体关系抽取获得了显著成果^[9]。

对知识图谱的研究,发源于 20 世纪的“引文网络”,目前知识图谱已运用到多个领域。陈强和代仕娅提出了一种将会计案防领域行业知识与大数据、知识图谱等技术相结合的智能化平台建设方案,能更高效准确地识别出可疑风险账户^[10]。杨波等人研究知识图谱在风险管理领域的应用,介绍了风险管理领域知识推理 (RMDKR) 的方法,主要包括传统的演绎推理方法、基于逻辑规则与案例的方法,以及基于人工智能的知识推理^[11]。知识图谱应用到食品监管领域方面的研究目前较少。张阿兰等利用 CiteSpace 工具进行关键词聚类,分析了 2003~2018 年中国知网上研究食品监管的论文,发现研究集中在“食品安全监管”“食品安全”“社会共治”等方面^[12]。秦丽等利用知识图谱 (KG) 技术提取了标准文件的内容与标准文件之间的引用关系,构建食品安全标准知识图谱,使食品安全标准与相关的食品生产过程可以建立联系^[13]。

2 技术路线

食品监管知识图谱构建过程主要包括两个步骤:实体识别、关系抽取。实体抽取部分通过构建实体识别模型 BiLSTM-CRF,训练模型并对抽取出的“主概念实体”进行聚类与归类;关系抽取部分先根据实体归类结果,确定同实体标签的实体间的分类关系,再构建“文本-实体”矩阵,分析包含某实体对的句子,确定不同实体之间的其他关系。构建流程如图 1 所示。

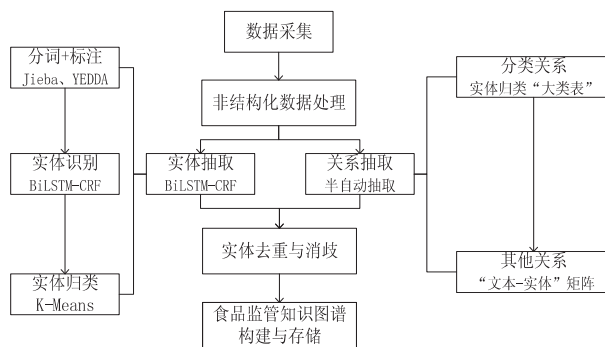


图1 食品监管知识图谱构建流程

3 基于 BiLSTM-CRF 模型的食品监管实体识别

首先,对采集的文本数据进行预处理,包括分词和词性标注,并确定实体标签;然后,基于 word2vec 向量化输入文本;接着,基于 BiLSTM-CRF 模型识别实体标签;最后,使用 K-means 聚类算法将相同实体归为一类。实体识别流程如图 2 所示。

3.1 数据预处理

采集的数据包括原国家食品质量监督检验中心发布的与食品监管相关的规章制度等文件、国家市场监

督管理总局官网 (<http://zwfw.samr.gov.cn>) 发布的食品监管政策公文等文件及百度百科相关词条等。共获取文本数据 45 篇。

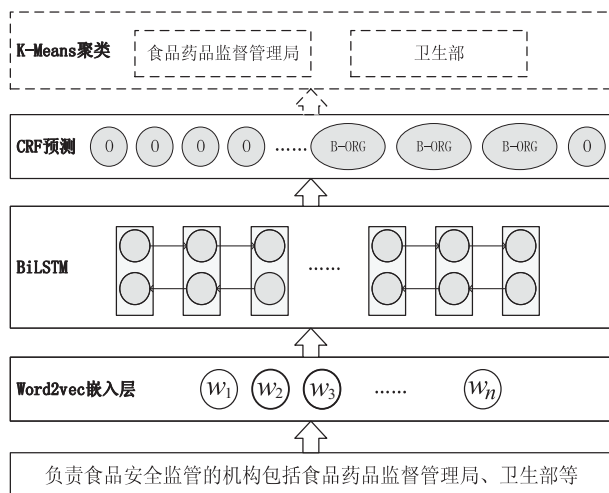


图 2 实体识别流程

处理的数据为中文文本,需要先使用 jieba 进行分词和词性标注,然后对分词结果进行词频统计,并结合对食品监管行业知识的考量,确定用于标注和识别的实体标签,如表 1 所示。标注后的训练数据共 292 917 行,按模型需求将其以句子为单位分行。

表 1 主概念图谱实体标签

标签	简称	描述
Food	FOD	食品(包括食品及原料)
Additive	ADT	添加剂(如防腐剂等)
Organization	ORG	机构(如食品监管机构)
Operation	OPE	食品监管工作(如抽检、行政处罚)
Content	CON	监管内容(如保质期、许可证)
Other	OTH	其他监管对象(如企业、个体工商户)

3.2 基于 BiLSTM-CRF 模型的食品监管实体识别

采用“自顶朝下”与“自底朝上”相结合的方式行实体识别,先确定实体标签,再用模型识别更多实体,并通过实体的聚类、对齐等工作进行优化。模型主要由输入层、网络层和输出层三部分组成。

输入层采用 Keras 的序列预处理。首先,根据训练数据词汇表 vocab 将输入文本转为字典 word2idx。然后,对文本转化的字典进行操作,获取字向量并通过 pad_sequences 函数进行序列填充和对齐。经过处理后,得到最终的输入向量,并传递给网络层。网络层的主要任务是对输入序列进行特征提取,也就是编码操作。该文采用双向的 LSTM 网络^[2,14],经过前向、后向传递,同时考虑上下文信息,处理模型输入的特征向

量,得到预测特征向量。

输出层是把网络层编码的输出进行最终处理,获得预测标签。虽然 BiLSTM 已经可以预测标签,但结果不够精确,只适合大规模但低精度的任务。而添加 CRF 层^[15]可以考虑相邻的输入样本,学习句子的约束条件,计算联合概率,从而优化整个序列的预测结果。

以准确率作为模型的性能指标,如下公式所示。

$$ACC = \frac{TP + TN}{n}$$

其中, n 表示样本总数, TP 表示正确判断为正向的样本数, TN 表示正确判断为负向的样本数。

将训练集与测试集传入模型进行训练,最终的训练集的 CRF 准确率(crf_accuracy)为 0.979 3,测试集 CRF 准确率(val_crf_accuracy)为 0.957 9。训练过程 10 个 epoch 的 CRF 准确率变化曲线如图 3 所示。

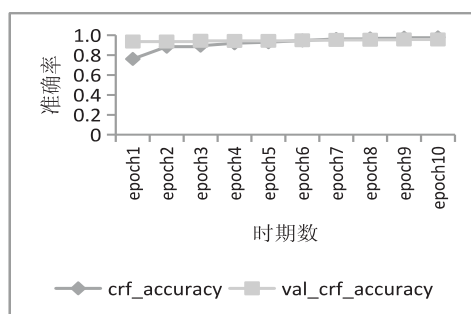


图 3 模型训练过程准确率曲线

句子“在中央一级,负责食品安全监管的机构包括食品药品监督管理局、卫生部、质检总局、国家工商总局、商务部等”的识别结果如图 4 所示,其中“食品药品监督管理局”“卫生部”“国家工商局”属于 ORG。



图 4 模型实体识别:单句识别

3.3 实体聚类对齐与存储

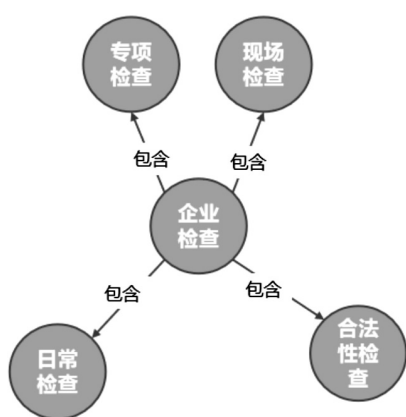
对模型识别出的实体进行人工选择,删去明显的错误实体,最终保留的实体数量为 333 个。去重后的实体中还存在一些同义实体,如“食品药品监督管理局”与“食品药品监督管理局”等,同时除同义词以外,对于实体中少量的实例数据,如“国家市场监督管理总局”“江苏省市场监督管理局”等,也可作为“同类词”用大类

“市场监管机构”来指代。因此,通过 K-Means 算法进行实体聚类, K 值设定为实体数量整除 5 再加 1,最后输出各标签内部的聚类结果。以“OTH”标签为例,该标签共 26 个实体, K 值为 6,最后共分为 6 个簇,如图 5 所示,可以看到“OTH_0”中以泛指性食品生产者为主,“OTH_2”以食品服务提供者为主。使用聚类后的实体构建“大类表”,最终得到“食品添加剂”“食品企业”等 68 个大类。

```
ent_cluster('OTH')

K: 6
OTH_0:
['食品生产经营企业', '食品生产企业', '食品摊贩', '特殊食品生产企业', '小作坊']
OTH_1:
['餐饮服务提供者', '网络餐饮服务第三方平台提供者', '网络食品交易第三方平台提供者']
OTH_2:
['生产企业']
OTH_3:
['生产经营者']
OTH_4:
['餐饮服务企业', '餐饮服务', '餐饮业', '糕点生产企业', '个体工商户', '酒生产企业', '农产品生产企业', '肉类协会', '肉制品生产企业', '生产经营企业', '食品加工业', '食品加工业协会', '食品企业', '食品行业协会', '特殊食品企业']
OTH_5:
['企业']
```

图 5 K-Means 实体聚类结果:以 OTH 为例



4 基于“文本-实体”矩阵的食品监管关系抽取

由于食品监督文本格式较为规范,使用监督学习效果不佳,因此,提出基于“文本-实体”矩阵的食品监管主概念关系抽取方法。首先,利用上文实体聚类获得的“大类表”;然后,构建以大类实体为列名、句子为行名的“文本-实体”矩阵;最后,将该对实体的关系归类为最接近的关系标签,实现关系抽取。

4.1 分类关系获取

分类关系即“上下位关系”,如“人民政府”的下位词是“中央人民政府”“地方人民政府”等。在“大类表”中,大类实体(class)与实体(name)自然呈现为分类关系,可以分别视为头实体与尾实体。该文共存储了 192 条“包含”关系。关系获取的部分结果如图 6 所示,“食品添加剂”包含“保鲜剂”“营养强化剂”等;“检验机构”包含“进出口检验机构”“复检机构”等;“企业检查”包含“专项检查”“现场检查”“日常检查”“合法性检查”等。

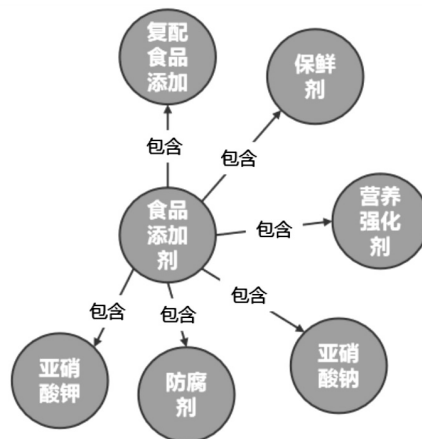


图 6 “包含”关系获取结果(分类关系)

4.2 基于“文本-实体”矩阵的关系获取

该文的关系抽取仅考虑“大类表”68 个大类间的关系。设大类 A 包含实体 a, 大类 B 包含实体 b, 则下位实体 a 与 b 的关系归为 A 与 B 的关系。以大类 ID 作为列名,待抽取文本作为行名,构建“文本-实体”矩阵,若某行文本包含某大类实体,则值为 1,否则为 0。

基于“文本-实体”矩阵,提出两种关系获取的方法:①对于重要实体,使用上文构建的“文本-实体”矩阵筛选出包含该实体的所有句子,以及同句的其他实体,获取该实体的所有关系;②对于其他实体,可以进

一步压缩“文本-实体”矩阵,仅保留句子与该句所含实体序列,重点关注不同标签的实体,批量获取不同实体对之间的关系。

(1) 指定实体的关系抽取。

对“文本-实体”矩阵,进行归类和人工判断,归纳实体间的关系,并保存为三元组。以“工商部门”大类 OR3 为例,可以看到与之相关的实体有“检验机构”OR6、“食品药品监管机构”OR12、“食品企业”OT1、“个体工商户”OT2,如图 7 所示。

text	OP8	OR1	OR2	OR3	OR4	OR5	OR6	OR7	OR8	OR9	OR10	OR11	OR12	OR13	OT1	OT2
在中央一	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
大多数省、	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
年月,由	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
国家工商	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
年,国家	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
从年月日	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
第一,食	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0

图 7 实体对及关系的确定:以大类 OR3 为例

观察包含图 7 OR3 的 7 行文本数据,发现 OR3 与 OR6、OR12 仅仅是被同时提到,而 (OR3, OT1) (OR3, OT2) 两个实体对关系较为密切,分别为 OR3“检查”OT1 与 OT2、OR3“排查”OT1 与 OT2;OR3“备案”OT1,如表 2 所示。

表 2 关系三元组示例(尚未去掉 id 列)

head	id1	rel	id2	tail
工商部门	OR3	检查	OT1	食品企业
工商部门	OR3	检查	OT2	个体工商户
工商部门	OR3	排查	OT1	食品企业
工商部门	OR3	排查	OT2	个体工商户
工商部门	OR3	备案	OT1	食品企业

text	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11
所使用的 A1	C1	C14	-	-	-	-	-	-	-	-	-
所使用的 A1	C1	C14	-	-	-	-	-	-	-	-	-
网络食品 A1	C11	C12	F1	OT3	OT4	-	-	-	-	-	-
网络食品 A1	C11	C12	F1	OT3	OT4	-	-	-	-	-	-
食品安全 A1	C13	-	-	-	-	-	-	-	-	-	-
致病性微 A1	C13	-	-	-	-	-	-	-	-	-	-
食品、食 A1	C13	-	-	-	-	-	-	-	-	-	-
致病性微 A1	C13	-	-	-	-	-	-	-	-	-	-
生产经营 A1	C13	-	-	-	-	-	-	-	-	-	-
进口商应 A1	C2	C3	C4	C11	C14	-	-	-	-	-	-
第九十八 A1	C2	C3	C4	C11	C14	-	-	-	-	-	-
食品添加 A1	C2	C3	C4	C5	C6	C9	C11	C14	-	-	-
第六十条 A1	C2	C3	C4	C5	C6	C9	C11	C14	-	-	-
食品添加 A1	C2	C3	C4	C6	C9	C11	C14	C15	OP3	-	-
第五十九 A1	C2	C3	C4	C6	C9	C11	C14	C15	OP3	-	-
第三十三 A1	C2	C3	C4	C9	C11	C14	-	-	-	-	-
食品生产 A1	C2	C3	C4	C9	C11	C14	F6	OT1	OT3	-	-
食品生产 A1	C2	C3	C4	C9	C11	C14	F6	OT1	OT3	-	-
进口的食 A1	C2	C9	C11	C14	F6	-	-	-	-	-	-
严厉打击 A1	C3	-	-	-	-	-	-	-	-	-	-
标注虚假 A1	C3	C4	-	-	-	-	-	-	-	-	-
食品和食 A1	C3	C4	-	-	-	-	-	-	-	-	-
生产经营 A1	C3	C4	-	-	-	-	-	-	-	-	-

图 8 “文本-实体”压缩矩阵示例

使用上述方法大致获取关系的同时,可以结合食品监管行业知识,归纳普遍关系,确定关系标签,使关系标准化。例如,表 2 的“排查”关系可以归为“检查”。确定标签的大致过程如下:分析 OTH 与 FOD、ADT 的关系,可以确定“生产”“采购”关系;筛选 ORG 与 OTH 实体,确定“检查”“许可”“认证”关系,以及 OTH 对 ORG 的“申诉”“备案”等其他工作统一用“监管”指代;筛选 ORG 与 FOD、ADT 实体,通过“检验机

(2)关系批量抽取。

在提取大量关系时,逐个筛选判断工作量较大,需要进一步压缩“文本-实体”矩阵,提取每个句子包含的实体序列。经统计,一行包含实体数最多为 11 个,因此,构建一个十二列的“文本-实体”压缩矩阵,将原本分散的实体压缩到前几列。该步骤虽丢失了实体的位置信息,不适合筛选单个实体的关系,但大大降低了关系批量提取的难度。

压缩后的“文本-实体”矩阵如图 8 所示,通过矩阵可以观察实体的分布状况,从而找到普遍关联的实体对及其关系,例如可以观察到“A1”与“C1”“C2”“C3”均有关系,即“食品添加剂需标注名称、地址、生产日期”,可以获取三个“包含”关系。

构”检验“特殊食品”等描述,确定“检验”关系;ORG 与 OPE 实体根据逻辑确定为“职责”关系;CON 作为监管内容,同时从属于监管工作与监管对象,考虑到关系可以通过标签区分,可以借用上文表达分类关系的“包含”关系。另外,机构还存在下级机构,若用“包含”易与分类关系混淆,所以确定“下辖”关系等。经过上述操作,确定了 11 个较为常见的关系标签,如表 3 所示。

表 3 主概念图谱关系标签

标签	关系	描述
生产	OTH—>FOD/ADT	食品生产企业生产食品、添加剂
采购	OTH—>FOD/ADT	食品企业采购食品、原料、添加剂
申诉	OTH/ORG—>ORG	企业或组织向上级机构申诉
检查	ORG—>OTH	机构与食品企业、个体工商户等的检查关系
许可	ORG—>OTH	监管机构与企业等的许可关系
认证	ORG—>OTH	监管机构与企业等的认证关系

续表 3

标签	关系	描述
监管	ORG→OTH/ORG	机构对企业及下级机构其他监管行为,如备案等
检验	ORG→FOD/ADT	检验机构与食品、添加剂的关系
职责	ORG→OPE	机构与其负责的监管工作的关系
下辖	ORG→ORG	机构与下辖机构

4.3 食品监管知识图谱构建

将实体及实体-关系表格导入至 Neo4j 中,生成食

品监管主概念图谱,如图 9 所示(仅展示部分)。圆圈表示实体,箭头表示实体之间的关系。

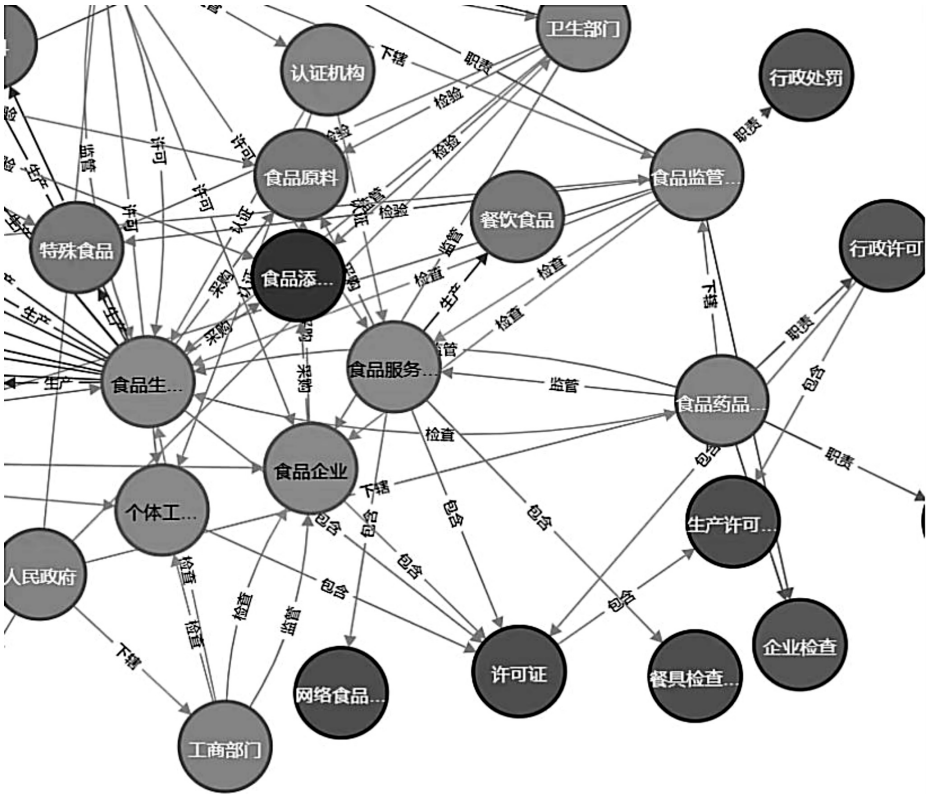


图 9 食品监管主概念图谱展示

5 结束语

研究了食品监管领域知识图谱的构建方法,通过探讨和实现实体及实体关系抽取算法,构建食品监管领域知识图谱。总结主要工作如下:(1)实现了 BiLSTM-CRF 实体识别模型的构建,利用 BiLSTM-CRF 模型对食品监管的政策公告等进行了主概念实体识别。包括从数据预处理开始的模型的训练与修改、实体的聚类归类与对齐等工作,模型的准确率为 0.957 9。(2)提出用构建“文本-实体”矩阵的方法实现关系抽取,在实体归类形成的 68 个大类的基础上,以大类实体为行,句子为列,判断每个句子是否包含某实体,形成一个 0-1 矩阵。据此可找出实体对及相关句子,归类到特定的关系。

参考文献:

[1] 邓攀. 食品流通领域的监管问题与对策[J]. 经济师,

2021(5):30-31.
[2] 尚福华,金泉,曹茂俊. 基于 Senna-BiLSTM-CRF 的测井实体抽取方法研究[J]. 计算机技术与发展,2021,31(12): 180-186.
[3] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv:1508.01991, 2015.
[4] 叶蕾,戴文华,钱涛,等. 基于 BiLSTM-CRF 的中文电子病历命名实体识别研究[J]. 湖北科技学院学报,2021, 41(2):83-89.
[5] 罗熹,夏先运,安莹,等. 结合多头自注意力机制与 BiLSTM-CRF 的中文临床实体识别[J]. 湖南大学学报:自然科学版,2021,48(4):45-55.
[6] HASHIMOTO K, MIWA M, TSURUOKA Y, et al. Simple customization of recursive neural networks for semantic relation classification[C]//Proceedings of the the 2013 conf on empirical methods in natural language processing. Seattle: [s. n.], 2013:1372-1376.
[7] ZENG D, LIU K, LAI S, et al. Relation classification via conv-
(下转第 214 页)