厨房饮食知识图谱的构建方法

袁琦　施银军　刘俊翔　俞贵涛　陈斌德
（宁波方太厨具有限公司，315336）

摘　要：提出一种厨房饮食领域（食材和菜谱为主）的知识图谱构建方法，通过自上而下的方式构建厨房饮食中食材和菜谱相关的概念（Schema）层，从数据中进行实体抽取来构建实例层，数据包括半结构化和非结构化两种类型数据。对于非结构化数据，采用深度学习 BiLSTM-CRF 模型来进行食材功效、食材搭配功效和菜谱功效的实体抽取。实验结果表明，与传统的机器学习算法 CRF 相比，该模型确实能有效地提高实体识别的 F 值。通过食材实体的别名关系来进行食材之间的实体链接，通过计算菜谱实体相似度来进行菜谱实体的链接。厨房饮食知识图谱采用原生图数据库 Neo4j 来存储。

关键词：厨房饮食知识图谱　BiLSTM-CRF　实体链接　图数据库

Method for Constructing Kitchen Diet Knowledge Graph

Yuan Qi，Shi Yinjun，Liu Junxiang，Yu Guitao，Chen Binde
（Ningbo Fotile Kitchenware Co.，Ltd.，315336）

Abstract: This paper proposes a knowledge graph construction method in the field of kitchen diet（mainly ingredients and recipes）. It constructs the concept（Schema）layer related to ingredients and recipes in the kitchen diet through a top-down manner，and implement entity extraction from the data to construct the instance level，the data includes two types of data，semi-structured and unstructured. For unstructured data，the deep learning BiLSTM-CRF model is used to perform entity extraction of ingredient efficacy，ingredient collocation efficacy，and recipe efficacy. Experimental results show that，compared with the traditional machine learning algorithm CRF，this model can effectively improve the F value of entity recognition. The entity links between the ingredients are made through the alias relationship of the ingredients，and the recipe entities are linked by calculating the similarity of the recipe entities. The kitchen diet knowledge graph is stored in the native graph database Neo4j.

Keywords: Kitchen diet knowledge graph，Bilstm-CRF，Entity link，Graph database

1　引言

随着中国经济社会的快速发展，人民生活水平的巨大提升，人们的饮食已经从吃得饱到吃得好，再到吃的健康转变，人们也越发重视饮食营养和膳食的平衡，对食材的功效、营养价值的了解，对食材搭配知识的了解，以及对菜谱功效和营养价值的了解是健康饮食的根本。同时食材和菜谱也有自己的适宜人群和禁忌人群，有些食材和菜适宜中老年人群，有些适宜有高血压的人群吃的，也有食材和菜谱是感冒的人不能吃的，也有些是肠胃不好的人忌食的。但是目前人们对这些知识都非常的缺乏，当人们想要了解这些厨房饮食养生知识的时候需要通过搜索引擎来搜索才能获取相关的网页和文档，并且需要自己进一步的阅读和筛选才能获得想要的知识，因此人们对能够直接获取厨房饮食知识的智能问答很有需要。工业界基于知识图谱的问答机器人有小米的小爱同学、微软小冰以及百度的小度等，因此构建厨房饮食知识图谱对实现饮食健康相关的智能问答有着很大的应用价值。

此外，随着生活质量的提高，人们也越来越喜欢通过电蒸箱、电烤箱等智能厨电工具进行烹饪和烘焙，来自己制作健康的美食，丰富自己休闲生活，这就需要我们的智能机器能够教人们制作健康美食。因此在电蒸箱和电烤箱中嵌入以厨房饮食知识图谱为基础的智

通讯作者简介：陈斌德，男，1982 年 3 月生，毕业于南京大学软件工程专业，现任宁波方太厨具有限公司智能研究院院长，主要从事物联网与人工智能等方向的研究，电子邮箱 chenbd@fotile.com。

能问答系统和智能菜谱是不可避免的趋势。目前很多垂直领域的公司都推出各自领域的知识图谱来改善服务质量，但是目前在智能厨电领域尚未出现成熟的能够基于厨房饮食知识图谱在厨电设备上直接回答人们关于食物营养和食物健康养生问题的厨电智能问答系统。同时厨房饮食知识图谱中的菜谱知识也是进行菜谱个性化推荐的基础，因此构建厨房饮食知识图谱具有重要意义。

2012年谷歌首次提出知识图谱（knowledge graph）[1]的概念，国外的通用领域知识图谱研究已经出了很多成果，例如国外的YAGO[2]、Freebase[3]等，这些通用知识图谱中包含了大量的常识知识，国内的包括Zhishi.me[4]等，是中文通用知识图谱。

在垂直领域方面，国内有宠物领域知识图谱的半自动化构建[5]，首先创建宠物知识图谱的概念层，通过CRF条件随机场结合宠物症状词典的方法进行症状命名实体识别，最后采用OrientDB图数据库进行宠物知识的存储。

领域知识图谱通常注重知识的层次结构，大部分垂直领域知识图谱都是通过半自动化的方式来构建。需要预先通过自上而下的方式构建领域概念层，然后再通过自下而上的方式构建领域数据层，同时运用多种实体抽取和关系抽取技术自动的抽取出置信度比较高的知识合并到领域知识图谱中。

目前实体抽取主要采用CRF[6]传统机器学习的方法，而采用基于深度学习模型BiLSTM-CRF的方法，与传统的机器学习算法CRF相比，能有效地提高实体识别的F值。

2 厨房饮食知识图谱的构建

厨房饮食知识图谱构建的流程如图1所示。

（1）概念层构建。通过对养生美食网站数据的分析从而自上而下的构建概念层（schema layer）。

（2）在半结构化和非结构化文本中进行抽取。从养生美食网站的半结构化文本中进行实体、属性和实体关系的抽取。从养生美食网站的非结构化文本中进行食材功效、食材搭配功效以及菜谱功效的命名实体识别和抽取。

（3）实体链接。通过食材实体的别名关系来进行食材之间的实体链接，通过计算菜谱实体相似度来进行菜谱实体的链接。

（4）知识存储。厨房饮食知识图谱使用Neo4j图数据库进行厨房饮食知识的存储。

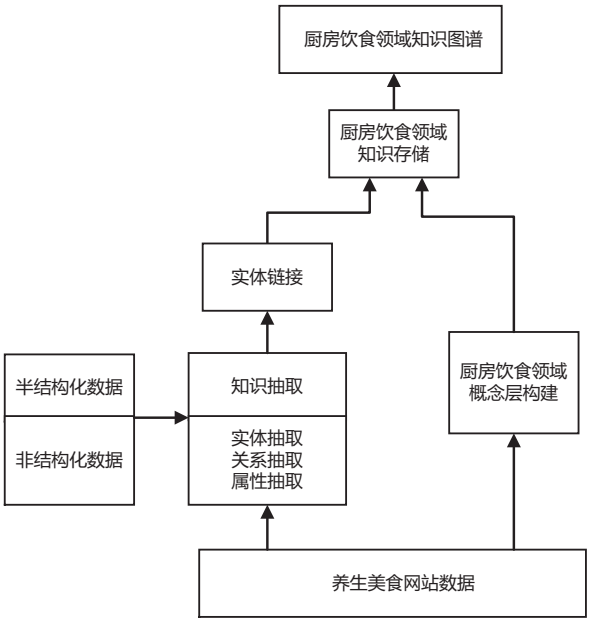

图1 厨房饮食知识图谱构建流程

2.1 厨房饮食概念层的定义和构建

厨房饮食知识图谱概念层的定义和构建是对整个厨房饮食知识图谱框架的构建，同时定义知识图谱中类以及类之间的关系。

本文设计和构建的是厨房饮食领域的知识图谱，构建了知识图谱的概念层，定义了六大类，包括食材品种、食材功效、食材搭配，搭配功效、菜谱和菜谱功效。其次是属性的定义：食材品种的属性包括食材名称、别名、概述、营养、适宜人群和禁忌人群；菜谱的属性包括菜谱名称、概述、适宜人群、禁忌人群、材料、做法步骤；食材功效的属性是功效；食材搭配的属性是搭配；搭配功效和菜谱功效的属性都是功效。

通过定义的六大类，创建了类之间5种语义关系，如下：

（1）e_HasCookTogether（有烹饪搭配）。食材品种——食材搭配，食材品种和食材搭配之间存在的关系。

（2）e_HasCookFunction（有搭配功效）。食材搭配——搭配功效，食材搭配和搭配功效之间存在的关系。

（3）e_HasFunction（有功效）。食材品种——食材功效，食材品种和食材功效之间存在的关系。

（4）e_HasRecipe（有菜谱）。食材品种——菜谱，食材品种和菜谱之间存在的关系。

（5）e_HasRecipeFunction（有菜谱功效）。菜谱——菜谱功效，菜谱和菜谱功效之间存在关系。

以上为厨房饮食知识图谱概念和语义关系的创建。厨房饮食知识图谱的概念层如图 2 所示。

2.2　数据源

厨房饮食知识图谱是从养生、美食等垂直网站上的数据中抽取知识，通过爬取垂直网站上的数据和百科知识，进行饮食实体、关系和属性的抽取，这些养生美食垂直网站也提供了质量比较高的领域半结构化数据。

2.3　从半结构化数据中抽取

本文主要从养生、美食网站上的半结构化数据中抽取食材品种、食材搭配和菜谱的实体、语义关系和实体属性。通过爬虫采集养生、美食网站的网页信息，然后进行清洗和解析。本文采用用来抓取网页的 urllib2 库和可以从 HTML 网页中解析数据的 beautiful soup 库，通过 urllib2+beautiful soup 抽取食材品种及食材品种属性、食材搭配及食材搭配属性、菜谱及菜谱属性的实体。在抽取实体过程中同时实现实体之间语义关系的挖掘，实现 5 种语义关系的获取。以食材红萝卜为例，如图 3 所示。

图 3 中，解析了食材红萝卜的网页抽取了红萝卜这个食材实体包括它的名称、别名、概述、营养价值、适宜人群和禁忌人群 6 个属性，根据对食材属性的定义，也就获取了 6 条属性值关系，在搭配食材文本中可以抽取红萝卜 + 菠菜、红萝卜 + 草鱼以及红萝卜 + 莴苣这 3 个食材搭配实体以及它们的搭配属性，获取了红萝卜食材品种实体以及红萝卜相关的食材搭配实体，也就获取了食材品种和食材搭配 e_HasCookTogether（有烹饪搭配）这种语义。图 3 中搭配食材的功效是一段非结构化文本，本文将采用 BiLSTM-CRF 模型来实现搭配功效的抽取，抽取出保持心血管的畅通、延缓衰老和强心健脾这 3 个搭配功效实体，同时也可以得到食材搭配和搭配功效 e_HasCookFunction（有搭配功效）这种语义。

2.4　从非结构化数据中抽取

2.4.1　数据集与标注

因为目前国内没有厨房饮食领域用于抽取食材、食材搭配以及菜谱的功效实体的开源数据集，所以本文自己构建训练和测试的语料，本文通过抽取食材功效为例来做实验，将 120 条食材描述的文本构建成训练集，40 条文本构建成测试集。当实验结果达到要求时，将用训练好的模型抽取非结构化文本中的食材功效、搭配功效以及菜谱功效的实体。

本文采用 BIO 对语料进行标注，标记为 B-FUNC，I-FUNC，O，分别表示功效实体的首部、功效实体的中部以及非功效实体内容。

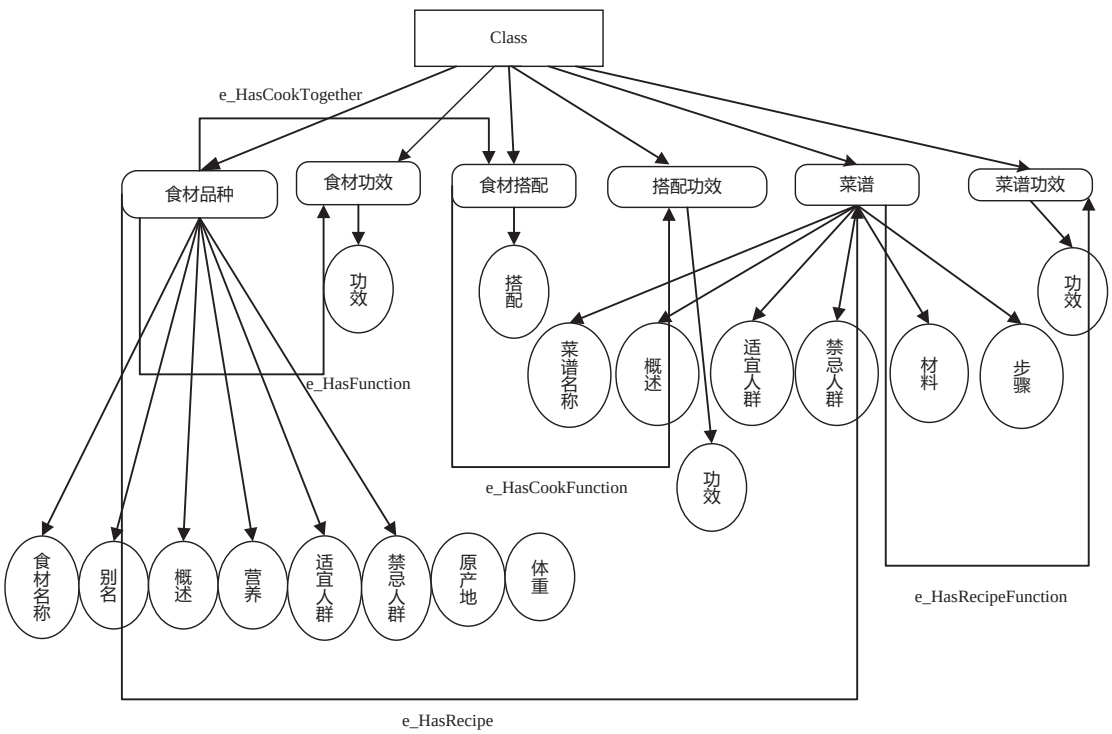

图 2　厨房饮食知识图谱概念层

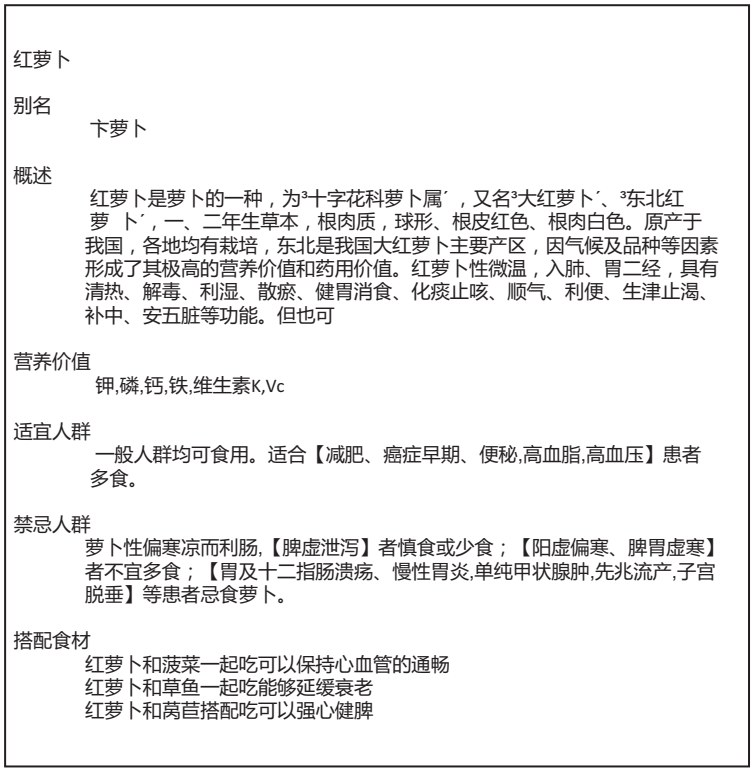

图 3　红萝卜食材

2.4.2　基于 BiLSTM-CRF 的命名实体识别方法

目前有很多算法可以进行，机器学习的常用算法有 CRF，本文用的是基于深度学习的命名实体模型 BiLSTM-CRF 命名实体识别模型[7]，BiLSTM-CRF 结合了双向长短时基于模型（BiLSTM Bidirectional Long Short-Term Memory）和条件随机场模型（CRF Conditional Random Fields）。条件随机场是一个序列标注算法，可以使用字、词语本身以及上下文特征，同时可以结合词典等特征，长短时记忆模型[8]通过引入输入门、遗忘门、存储单元以及输出门的控制机制，能够很好的提高模型利用长距离历史信息的能力，是一种特殊的循环神经网络。

BiLSTM-CRF 模型既能使用双向长短时记忆网络 BiLSTM 提取文本信息的特征，又可以利用条件随机场 CRF 自动学习一些约束性规则，更好的利用上下文信息。模型整体构造图 4 所示。

第一层是输入层，输入层以字为单位进行输入，look-up layer 将初始输入经过预训练的 Embedding 向量矩阵映射成序列 $X=(x_1, x_2, x_3, \cdots, x_4)$。look-up layer 层输出的字向量序列 X 是 BiLSTM 层每个时间点 t 的初始输入值，BiLSTM 层是隐含层，获得字的内在特征。在 BiLSTM 层中，输入向量的顺序序列是前向 LSTM 层的输入，而输入向量序列的逆序序列是后向 LSTM 层的输入。之后 BiLSTM-CRF 模型将正向 LSTM 层在 t 时刻输出的隐状态序列和反向 LSTM 层在 t 时刻输出的隐状态序列进行拼接，从而得到完整的隐状态序列 $C=(c_1, c_2, c_3, \cdots, c_n)$。之后通过一个线性层将隐状态向量映射到 k 维，k 是数据集的标签总数，映射后得到 $P=(p_1, p_2, p_3, \cdots, p_n)$，$P$ 是 $n \times k$ 的矩阵输入到 CRF 层中。

CRF 层是序列标注层，对整个句子级层面的序列进行标注，通过学习出一些约束性规则，提高预测的准确率。

对于输入一个句子 $X=(x_1, x_2, x_3, \cdots, x_n)$，会得到预测序列 $y=(y_1, y_2, y_3, \cdots, y_n)$，那么它的得分为：

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \quad (1)$$

A 是状态转移矩阵，$A_{y_i, y_{i+1}}$ 是 y_i 到 y_{i+1} 转移概率，p_{i,y_i} 表示 BiLSTM 层在第 i 个位置输出标签为 y_i 的概率，整个句子得分为各个位置的分数总和。CRF 层之后接入 softmax 函数实现分数的归一化，序列 y 的概率为：

$$p(y \mid X) = \frac{e^{s(X, y)}}{\sum_{\tilde{y} \in Y_X} e^{s(X, \tilde{y})}} \quad (2)$$

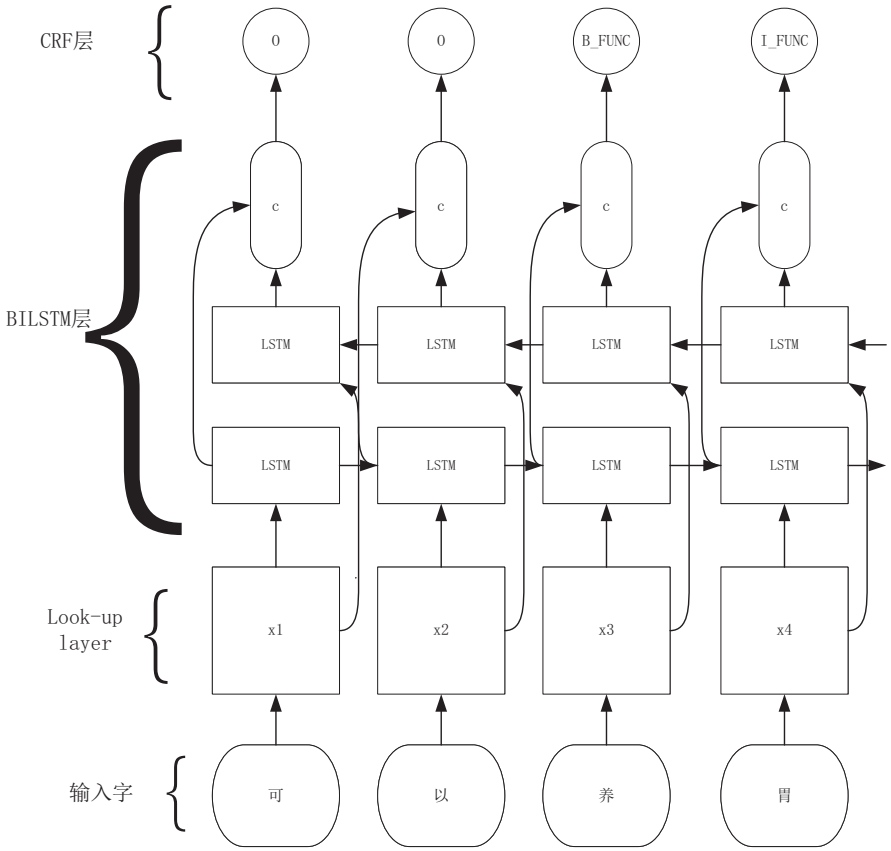

图 4　BiLSTM-CRF 模型结构

在训练中，我们需要最大化概率$p(y \mid X)$，利用梯度下降法，最大化对数似然函数：

$$\log(p(y \mid X)) = s(X, y) - \log(\sum_{\tilde{y} \in Y_x} e^{s(X, \tilde{y})}) \quad (3)$$

$$= s(X, y) - \log(\sum_{\tilde{y} \in Y_x} s(X, \tilde{y}))$$

在预测时，根据训练好的模型参数对序列进行预测，预测最优结果为y^*

$$y^* = \underset{\tilde{y} \in Y_x}{\operatorname{argmax}}(s(X, \tilde{y})) \quad (4)$$

2.4.3　实验与结果

本文使用标注的 160 条数据集做命名实体识别试验，其中 120 条文本是训练集，40 条文本是测试集。试验采用的评价指标为 Precision（精确率，P）、Recall（召回率，R）和 F 值（F1），公式如下：

$$P = \frac{\text{正确识别出的实体个数}}{\text{识别出的实体个数}} \times 100\% \quad (5)$$

$$R = \frac{\text{正确识别出的实体个数}}{\text{标准结果中的实体个数}} \times 100\% \quad (6)$$

$$F = \frac{2PR}{P + R} \times 100\% \quad (7)$$

进行食材功效实体抽取实验的平台环境为：操作系　统 Ubuntu 18.04，CPU 为 Intel（R）Xeon（R）Gold 6130 ，RAM 256G，显卡为 NVIDIA Tesla P100，16G。

CRF 模型和 BiLSTM-CRF 模型进行了对比实验，通过两个实验看食材功效实体抽取的效果，对比实验结果如表 1 所示：

表 1　对比实验结果

方法	Precision	Recall	F1
CRF	0.8907	0.8897	0.8902
BiLSTM-CRF	0.9039	0.9035	0.9037

由表 1 可知，BiLSTM-CRF 模型比 CRF 模型在精确率上提高了 1.32%，在召回率上提高了 1.38%，在 F 值上提高了 1.35%，这是在数据比较少的情况下的实验结果，已经有了一些提升，如果在数据比较多的情况下，深度学习模型 BiLSTM-CRF 的实体抽取的优势将更加明显。实验表明，与传统的机器学习算法 CRF 相比，BiLSTM-CRF 模型确实能有效地提高实体抽取的 F 值。

2.5 实体链接

实体链接的核心是计算实体之间的语义相似度，针对厨房饮食领域我们获取到的关于食材和菜谱的相关知识，对于食材实体我们用食材别名关系进行实体链接，这样我们能达到食材链接的百分之百的准确率，例如我们获取到了包菜食材的别名有包心菜、卷心菜、洋白菜和圆白菜等，我们将这些食材实体通过食材别名关系进行链接。关于菜谱实体我们用编辑距离的算法进行实体相似度的计算来进行菜谱实体的链接，例如枸杞猪心汤和猪心枸杞汤这两个菜谱实体其实是同一个菜谱，通过用编辑距离计算枸杞猪心汤和猪心枸杞汤这两个实体的相似度，就可以将这两个菜谱实体进行链接。

2.6 知识存储

本文使用的是图数据库Neo4j，是一个高性能的NoSQL[9]原生图数据库。它具备成熟数据库所有的特性，使用图相关的概念来描述数据模型，Neo4j创建的图用顶点和边构建一个有向图，可以很直接的使用图中节点和边的关系来建模。所以Neo4j图数据库存储数据要优于关系数据库，使用Cypher[10]语言作为查询语言，Neo4j已经成为主流的用来存储知识图谱的图数据库。

根据概念层的定义，创建概念类包括食材品种（v_Ingredient）、食材搭配（v_CookTogether）、食材功效（v_FoodFunction）、菜谱（v_Recipe）、菜谱功效（v_RecipeFunction）、搭配功效（v_CookFunction）、有烹饪搭配（e_HasCookTogether）、有搭配功效（e_HasCookFunction）、有功效（e_HasFunction）、有菜谱（e_HasRecipe）和有菜谱功效（e_HasRecipeFunction）。

在创建好厨房饮食知识图谱的概念层之后，导入对应标签中的所有实体以及实体之间的关系信息，同时使用Cypher查询语句防止实体与关系的重复导入。

3 结论

本文研究了一种厨房饮食领域基于深度学习的知识图谱构建方法。首先构建厨房饮食的概念层，也就是对整个厨房饮食知识图谱框架进行了构建，然后从养生、美食网站提供的半结构化文本中抽取厨房饮食领域的相关实体和关系，从非结构化文本中，用基于深度学习的BiLSTM-CRF命名实体模型进行食材功效、搭配功效以及菜谱功效的实体识别和获取，相比较于传统的CRF模型，该模型能有效提升命名实体识别的效果。同时对食材实体通过别名关系进行实体链接，对菜谱实体通过编辑距离进行实体相似度计算实现实体链接。最后用原生图数据库Neo4j进行知识的存储。

构建厨房饮食知识图谱是一项长期性的工作，还有很多需要改进的地方，比如知识推理，根据目前厨房饮食知识图谱中已有的知识，推理和补全出已存在厨房饮食知识图谱中但未被发现的知识。建立厨房饮食知识图谱的更新机制，不断提高知识图谱的实体数量和质量，让厨房饮食知识更加丰富。

总的来说，本文设计并实现了基于深度学习的厨房饮食知识图谱，该知识图谱为厨房饮食领域的智能应用提供了语料基础，同时该知识图谱为厨房饮食领域的智能问答机器人奠定了基础，具有很重大的意义。

参考文献

[1] SINGHAL A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012, 16.

[2] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial InTelligence, 2013, 194: 28-61.

[3] Yue Bin, Gui Min, Guo Jiahui, et al. An effective framework for question answering over freebase via reconstructing natural sequences[C]// Proc of International Conference on World Wide Web Companion: International World Wide Web Conferences Steering Committee. New York: ACM Press, 2017: 865-866.

[4] Niu Xing, Sun Xinruo, Wang Haofen, et al. Zhishi. me-weaving Chinese linking open data[C]// Proc of International Semantic Web Conference. Berlin: Springer, 2011: 205-220.

[5] 袁琦，刘渊，谢振平，陆菁. 宠物知识图谱的半自动化构建方法[J]. 计算机应用研究，2020，37(01)：178-182.

[6] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proc of the 18th International Conference on Machine Learning. Burlington: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[7] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition[J]. 2016: 260-270

[8] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[9] Ahmadian M, Plochan F, Roessler Z, et al. Secure NoSQL: an approach for secure search of encrypted nosql databases in the public cloud[J]. International Journal of Information Management, 2017, 37(2): 63-74.

[10] Panzarino, Onofrio. Learning cypher write powerful and efficient queries for Neo4j with Cypher its official query language[J]. 2014.