

用 BERT 和改进 PCNN 模型抽取食品安全领域关系

赵良^{1,2}, 张赵玥³, 廖子逸⁴, 王玲^{1,2}

(1. 华中农业大学信息学院, 武汉 430070; 2. 湖北省农业大数据工程技术研究中心(华中农业大学), 武汉 430070; 3. 华中科技大学网络安全学院, 武汉 430074; 4. 华中科技大学武汉国家光电研究中心, 武汉 430074)

摘要: 为了提高食品安全领域关系抽取的效率和准确性, 该研究在收集食品安全领域语料的基础上, 对语料中相应的实体和关系进行标注, 构建可用于食品安全领域关系抽取的专业数据集。同时, 提出面向食品安全领域的基于 BERT-PCNN-ATT-Jieba 的关系抽取模型, 该模型使用基于转换器的双向编码器表征量 (Bidirectional Encoder Representations from Transformers, BERT) 预训练模型生成输入词向量, 并结合分段卷积神经网络 (Piecewise Convolutional Neural Network, PCNN) 模型的分段最大池化层能极大程度捕获句子局部信息的特点, 在分段最大池化层与分类层之间添加了注意力机制, 以进一步提取高层语义。此外, 考虑中文语料的特性, 在 BERT 模型进行随机掩码切分之前, 采用 Jieba 分词技术对中文语料进行分词, PCNN 模型在执行掩码语言模型 (Masked Language Model, MLM) 时以词为单位进行掩码, 使得输入到训练模型中的句子尽可能减少语义损失, 以实现更高效的关系抽取。在该研究构建的数据集基础上, 将 BERT-PCNN-ATT-Jieba 模型与经典的卷积神经网络 (Convolutional Neural Network, CNN)、PCNN 模型、以及结合 BERT 的 CNN、PCNN、PCNN-ATT、PCNN-Jieba 等 6 个模型进行比较, 该研究提出的 BERT-PCNN-ATT-Jieba 模型取得更优的性能, 其准确率达到 84.72%, 召回率达到 81.78%, F_1 值达到 83.22%。该模型为食品安全领域的知识抽取提供参考, 为该领域知识图谱的自动化构建节约了成本, 同时为基于该领域知识图谱的知识问答、知识检索、数据共享及食品安全智慧监管等应用提供依据。

关键词: 食品安全; 模型; 关系抽取; 知识图谱; 注意力机制; BERT; PCNN

doi: 10.11975/j.issn.1002-6819.2022.08.030

中图分类号: TP391

文献标志码: A

文章编号: 1002-6819(2022)-08-0263-08

赵良, 张赵玥, 廖子逸, 等. 用 BERT 和改进 PCNN 模型抽取食品安全领域关系[J]. 农业工程学报, 2022, 38(8): 263-270.

doi: 10.11975/j.issn.1002-6819.2022.08.030 <http://www.tcsae.org>

Zhao Liang, Zhang Zhaoyue, Liao Ziyi, et al. Relationship extraction in the field of food safety based on BERT and improved PCNN model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(8): 263-270.

(in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.2022.08.030 <http://www.tcsae.org>

0 引言

食品安全关乎每个人的健康, 但目前食品安全大多存在难追溯、难控制、难防范的特点。食品安全领域的的数据以食品为中心, 包括食品原材料在种养殖过程中农药、兽药使用及残留数据, 食品在生产、储存过程中加入的添加剂、真菌毒素及其他化学、物理危害物相关数据, 食品本身微量元素数据, 食品与疾病健康对应关系的数据等。食品安全数据的应用价值, 主要在个人健康、食品研究、疾病预测、企业管理以及食品问题监管等方面。为了发现食品安全全链条不同部分数据之间的联系, 解决食品安全数据共享、食品安全智慧监管以及食品全链条溯源等问题, 构建面向食品安全领域的知识图谱具有重要意义。

在谷歌公司正式提出知识图谱 (Knowledge Graph, KG) 概念^[1]之后, 基于人工智能和自然语言处理技术的

发展, 知识图谱的出现成为连接两者的桥梁。知识图谱能够从海量无规则数据中抽取结构化信息^[2], 从而描述生活中存在的实体特征以及实体之间的关系^[3]。知识图谱按照内容主要分为开放领域知识图谱^[4]和垂直领域知识图谱^[5]。按照技术主要分为信息抽取、知识融合、知识加工、图谱应用等^[6]。信息抽取作为连接大规模数据集和知识图谱应用系统的重要媒介, 自动化高且准确的抽取手段则显得尤为重要。信息抽取主要包括 3 个子项任务: 实体抽取、关系抽取和事件抽取^[7]。而关系抽取作为信息抽取与信息检索等领域的核心任务和重要环节, 能够从文本中抽取实体之间或者实体与属性之间的语义关系, 例如得到 (实体, 关系, 实体) 或者 (实体, 属性, 属性值) 三元组^[8-9]。

随着机器学习以及知识图谱技术的发展, 关系抽取的算法层出不穷。从研究算法上看, 目前关系抽取技术主要分为 3 种: 第一种是基于手写规则^[10]的关系抽取方法。如 Aitken^[11]在一篇研究全球变暖的文章中, 将 371 条句子作为训练集, 定量定性谓词并进行提取, 准确率达到 66%。该方法的缺点是只能适用特定数据集, 三元组的查准率较高, 查全率比较低, 不能准确地查找所有满足要求的三元组。第二种是基于传统机器学习的关系

收稿日期: 2021-10-20 修订日期: 2022-03-29

基金项目: 国家重点研发计划项目 (2018YFC1604005); 中央高校基本科研业务费专项资金资助 (2662019PY070, 2662022JC004, 2662022XXYJ001)
作者简介: 赵良, 博士, 副教授, 研究方向为农业大数据、知识图谱等方面。
Email: zhaoliang323@mail.hzau.edu.cn

抽取方法。主要包括基于特征向量^[12]的方法, 基于核函数^[13]的方法以及条件随机场 (Conditional Random Field, CRF)^[14]的方法。如王东波等^[15]结合情报学数据获取、标注和组织的方法, 运用 CRF 机器学习模型, 在标注 15 000 字的语料库基础上进行实体抽取预测, F_1 值达到 91.94%。该方法的缺点是容易造成欠拟合问题且参数调整过程较为复杂。第三种是基于深度学习^[16]的关系抽取方法。如 Zeng 等^[17]在特征向量和核函数基础上提出卷积神经网络 (Convolutional Neural Network, CNN) 模型。王庆棒等^[18]基于 CNN 和双向长短记忆网络 (Bi-directional Long Short-Term Memory, BiLSTM) 模型对于食品舆情实体关系抽取, 准确率提升至 80%。Zeng 等^[19]将分段卷积神经网络 (Piecewise Convolutional Neural Network, PCNN) 模型应用于关系抽取中, 平均准确率提升至 78.3%。武小平等^[20]利用来自转换器的双向编码器表征量 (Bidirectional Encoder Representation from Transformers, BERT) 模型与 CNN 模型结合, 使得 F_1 值提升至 83%。

本文在尚无食品安全领域公开数据集的情况下, 收集权威机构的语料信息, 同时在专家指导下完成实体和关系的标注, 构建食品安全领域数据集。针对 PCNN 模型生成词向量不够精准以及中文以词为粒度分割的问题, 本文在构建的食品安全领域数据集基础上, 利用 BERT 和 PCNN 模型, 并结合分词技术、注意力机制等方法完成食品安全领域数据的关系抽取。并将改进的方法与目前主流的 CNN、PCNN、BERT-PCNN 等模型进行对比, 以期三元组自动化抽取提供参考。本文对食品安全领域相关数据的关系抽取进行研究, 以此来提高该领域内知识抽取的效率和准确性, 为自动化构建领域知识图谱提供参考, 而知识图谱可为食品安全数据共享、食品安全智慧监管和食品全链条溯源等问题提供底层知识库基础。

1 食品安全领域数据集构建

1.1 食品安全领域数据获取

本文在食品安全相关数据集鲜有公开的情况下, 通过 Scrapy 框架^[21]爬取百度百科 (<https://baike.baidu.com/>)、食品伙伴网 (<http://www.foodmate.net/>)、国家市场监督管理总局官网 (<https://www.samr.gov.cn/>) 等数据, 并由专家进行整理, 选择语义完整且丰富的语料, 构建相对完善的数据集, 数据源分布如表 1 所示。其中源数据格式为爬取之后的数据原始格式; 处理数据格式为文本预处理后的格式, 即如果是表格则直接提取三元组, 如果是文本则标注三元组位置。

表 1 数据源分布

Table 1 Distribution of data source

源数据格式 Source data format	处理数据格式 Processed data format
附录表格	实体+属性+属性值
正文文本	句子+关系+实体 1+实体 1 位置+实体 2+实体 2 位置

1.2 食品安全数据分类

结合食品安全专家建议以及知识推理过程, 对食品从种养殖、仓储、加工以及人类疾病等数据进行处理、存储及标注不同的实体。实体的分类描述如表 2 所示。

表 2 实体分类描述

Table 2 Description of entity classification

实体类别 Entity class	实体属性 Entity properties	实体举例 Entity example
食品	食品名称, 食品分类	小麦, 谷类
国家标准	标准名称, 标准编号	食品添加剂使用标准 GB2713-2015
食品营养值	营养名称, 营养值	蛋白质, 336 kJ
人体部位	人体部位名称	眼睛, 皮肤
农药	农药名称, 限量值等	苯乙酸, 2.5 mg·kg ⁻¹
化合物	化合物名称, 限量值等	三氯乙烯, 70 g
疾病	疾病名称	癫痫, 败血症
不良反应	不良反应名称	呕吐, 头晕
性别	男, 女	男性, 女性
年龄	具体年龄, 年龄阶段	10 岁, 青年

根据上述语料库和实体类别, 构建食品安全关抽取数据集。分类完成之后的数据集关系定义描述如表 3 所示。

表 3 关系定义描述

Table 3 Description of relationship definition

头实体分类 Head entity type	尾实体分类 Tail entity type	关系 Relation	描述 Description	数目 Number
食品	食品	包含	食品之间的关系	1 415
化合物或农药	化合物或农药	属于	化合物之间的关系	1 189
疾病	疾病	部分	疾病之间的关系	1 357
农药	症状	导致	农药超标可能引起的症状关系	1 299
农药	部位	损害	农药超标可能损害人体的部位关系	792
疾病	不良反应	症状	疾病可能导致的不良反应关系	1 281
疾病	性别或年龄	易感人群	发病率高的性别或者年龄关系	1 240

在整理的 8 573 条语料中, 标注了 7 类关系, 且各关系的数目较为平均。将整理好的数据集依据不同关系按照 8 : 2 比例划分成训练集和测试集。部分标注样本样例如表 4 所示。与同类数据集相比, 该数据集充分涵盖了食品安全各个阶段的数据, 同时也能够部分解决语料中实体重叠或者含有多个三元组的抽取问题。如语句 2) 可以抽取 2 个“属于”关系的三元组, 语句 3) 可以抽取“部分”、“易感人群” 2 个不同关系的三元组。

1.3 食品安全数据存储

对食品安全数据采取图的方式建模及存储。本文使用 neo4j 数据库^[22]存放实体关系三元组。知识图谱的节点为实体, 边为关系, 并由头实体指向尾实体, 不同实体的互联形成一张巨大的食品安全知识图谱网。通过 Cypher 语言^[23]将收集的三元组导入图数据库中, 并对数据进行管理。neo4j 数据库中部分数据展示如图 1 所示。

表 4 部分语料描述
Table 4 Description of partial corpus

语料 Corpus	头实体 Head entity	尾实体 Tail entity	关系种类 Kind of relationship
1) 黄瓜是常见的一种蔬菜，味道清香，适量的吃一些黄瓜可以起到促进肠道蠕动的作用。	黄瓜	蔬菜	包含
2) 多粘菌素 B 是多粘菌素的一种，由两种非常近似的化合物混合而成：多粘菌素 B1 和多粘菌素 B2。	多粘菌素 B1 多粘菌素 B2	多粘菌素 B 多粘菌素 B	属于
3) 恶性淋巴瘤又称“淋巴瘤”，是原发于淋巴结或其他淋巴组织的恶性肿瘤，是我国常见的十大恶性肿瘤之一。本病多见于中、青年，男性患者多于女性。本病按其细胞成分的不同可分为何杰金氏病和非何杰金氏淋巴瘤两大类。	非何杰金氏淋巴瘤 恶性淋巴瘤	恶性淋巴瘤 男性	部分 易感人群
4) 低浓度酚类可引起蓄积性慢性中毒，高浓度酚类可引起急性中毒以致昏迷死亡。酚类慢性中毒表现为头晕、腹泻等症状。	低浓度酚	腹泻	导致
5) 吡螨胺吞咽会中毒。可能导致皮肤过敏反应，吸入有害，长期或反复接触可能对器官造成伤害。	吡螨胺	皮肤	损害
6) 卡氏肺囊虫病是卡氏肺囊虫感染所引起的一种原虫病，临床主要表现为干咳、呼吸困难和发绀等。其病原特征为间质性肺炎，故称卡氏肺囊虫病肺炎。	卡氏肺囊虫病	干咳	症状

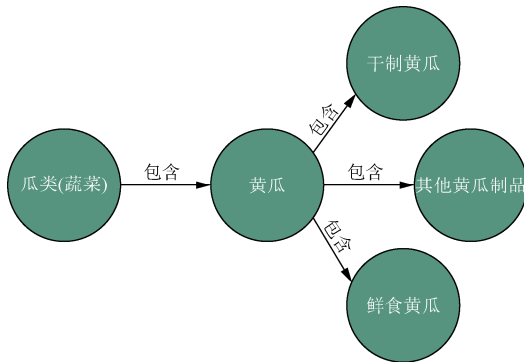


图 1 neo4j 数据库部分数据示例图
Fig.1 Partial data sample of neo4j database

2 食品安全领域关系模型

关系抽取流程如图 2 所示。首先对大规模文本采用手工标注的方法进行预处理，之后将得到的语料库利用 BERT 模型等技术得到句子的词嵌入向量和位置嵌入向量，将词嵌入向量与位置嵌入向量拼接得到句子的向量表示，向量表示的结果输入到 PCNN 神经网络模型进行训练。最后对测试集的句子进行测试，通过比较相应评估指标综合评价模型性能。

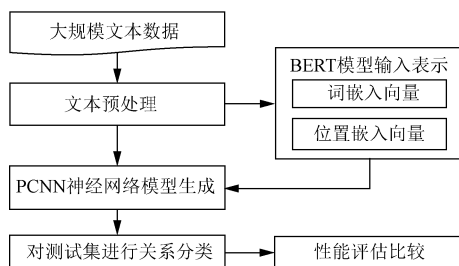


图 2 食品安全关系抽取流程图
Fig.2 Flow chart of food safety relationship extraction

2.1 BERT 向量化

随着自然语言和深度学习的发展，向量模型主要包括 one-hot^[24]、word2vec^[25]、语言模型的词向量 (Embeddings from Language Models, ELMo^[26])、生成式预训练模型 (Generative Pre-Training, GPT^[27]) 以及基于转换器的双向编码器表征量 (Bidirectional Encoder Representation from Transformers, BERT^[28]) 等。one-hot 的缺点是会导致特征向量空间非常大，不利于机器的存储与计算；word2vec 的缺点是无法区分一词多义问题；ELMo 的缺点是特征提取能力较弱^[29]；GPT 的缺点是无法准确利用词语上下文的关联信息；而 BERT 模型是结合 ELMo 与 GPT 模型优势的新型向量模型，该模型生成的词向量可以捕获更多的上下文文本特征^[30-32]。本文利用 BERT 模型生成词向量和位置向量，并将得到的词向量和位置向量进行拼接后，得到每个句子的输入表示。

2.1.1 词向量表示

对于一个由 m 个单词组成的句子 $\text{sentence} = \{\text{word}_1, \text{word}_2, \text{word}_3, \dots, \text{word}_m\}$ ，将每一个单词映射到一个固定长度的向量，该向量表示该词的语义关系，其中在句子开头有一个特殊符号 [CLS]，多句分隔或句子结束后有一个特殊符号 [SEP]。词向量表示如图 3 所示。



图 3 词向量表示描述图

Fig.3 Description of word vector representation

2.1.2 位置向量表示

位置向量的生成是以头实体和尾实体为基准，分别对于句子中的每个词求出距头实体和尾实体的相对位置，存储在距离头实体的位置向量 (head_pos) 和距离尾实体的位置向量 (tail_pos) 中。位置向量生成流程如图 4 所示。

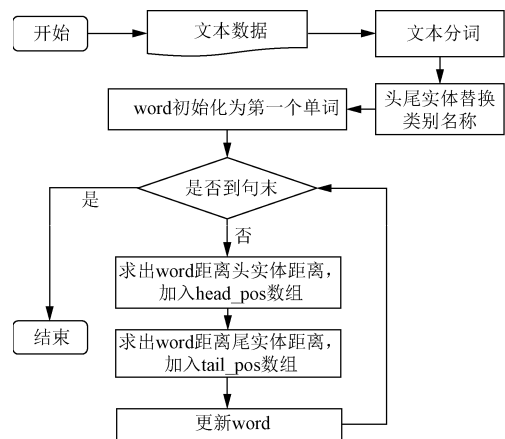


图 4 位置向量生成流程图

Fig.4 Flow chart of position vector generation

位置向量生成示例如图 5 所示。按照上述规则进行计算，得到 head_pos 结果为：

[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19]
tail_pos 结果为:
[-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]
head_pos 和 tail_pos 共同构成整句话的词向量。

小麦是三大谷物之一，几乎全食用，仅约有六分之一作为饲料使用。

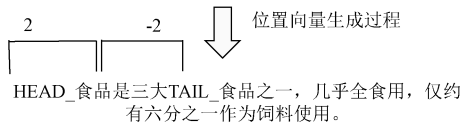
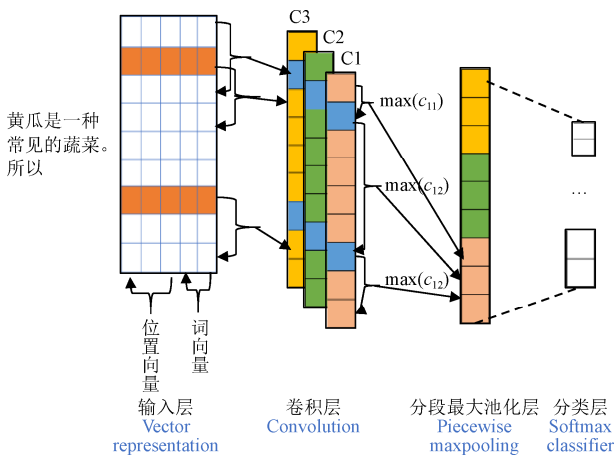


图 5 位置向量生成示例图

Fig.5 Example of position vector generation

2.2 PCNN 模型构建

分段卷积神经网络 (Piecewise Convolutional Neural Network, PCNN) 处理过程与卷积神经网络 (Convolutional Neural Network, CNN) 相似，同样适用于处理大规模文本数据，被广泛应用在自然语言处理、声音、图像等方面。PCNN 模型如图 6 所示。该模型包含输入层、卷积层、分段最大池化层、分类层。PCNN 模型与 CNN 模型区别在于用分段最大池化层替换原有的池化层，其中输入层的向量通过 BERT 模型计算得到。与传统的 PCNN 模型相比，本文构建的 PCNN 模型的改进有：1) 改变了原有输入表示的计算方法；2) 在卷积层中采用多尺寸卷积核计算方式关注语料库多样的信息，以及为解决神经元坏死问题采用的 GELU 函数；3) 在分类层中采用 Adam 算法^[33]自动计算一阶矩和二阶矩，从而动态更新每个参数的学习率。



注：C1、C2、C3 均为卷积层的一层。
Note: C1, C2 and C3 are all one layer of convolution layer.

图 6 PCNN 模型结构图

Fig.6 PCNN model structure

2.2.1 卷积层

卷积层的本质是向量的运算，卷积的目的是根据句子上下文中词语的语义特征拼接而得到整个句子的特征。卷积层通过高层特征限制输入层与隐藏层之间连接元的数目，从而减少模型训练的参数。多尺寸卷积核关注于不同特征值，使得训练模型可以全面的分析语句信息。综上，本研究中卷积层数设置为 3，每个卷积包含 100 个卷积核，大小分别为 1×3，1×5，1×9。

根据食品安全领域数据的特点，本研究使用 GELU 函数作为激活函数，GELU 函数结合了非线性计算与随机正则化计算，相较于 RELU 函数能解决神经元可能坏死的问题。GELU 函数计算公式为：

$$GELU(x) = \frac{1}{2}x \left(1 + \tanh \left[\sqrt{\frac{2}{\pi}} (x + 0.044715x^3) \right] \right) \quad (1)$$

式中 x 是具有零均值和单位方差的高斯随机变量， \tanh 为双曲正切函数。

2.2.2 分段最大池化层

分段最大池化层与最大池化层相比，在捕捉更高层语义特征的同时还能利用实体与实体之间的位置关系，具体过程如图 7 所示。分段最大池化层根据头实体和尾实体的位置将句子分为三部分：第一部分是句首到第一个实体 (包含第一个实体)，第二部分是两个实体之间 (包含头尾两个实体)，第三部分是另一个实体到句尾 (包含另一个实体)。对分段之后的句子进行填充像素 (padding)，padding 的目的是方便计算，处理过程以最长的部分为准，对另外两部分进行填充，如果该位置存在元素设置为 0，不存在设置为 1。

对于一个由 m 个单词组成的句子 $sentence = \{word_1, word_2, word_3, \dots, word_m\}$ ，根据头尾实体分为 3 个片段 $\{c_1, c_2, c_3\}$ ，经分段池化层处理后的输出向量为：

$$\begin{cases} P_i = \{P_{i1}, P_{i2}, P_{i3}\} \\ P_{ij} = \max(c_{ij}) \end{cases} 1 \leq i \leq m, 1 \leq j \leq 3 \quad (2)$$

式中 n 表示卷积核的个数， P_i 表示第 i 个句子的结果，拼接所有卷积核得到分段池化层 $P_i : n$ ， c_{ij} 为第 i 个句子中第 j 个分段， P_{ij} 为第 i 个句子中第 j 个分段的结果。经非线性函数输出 g 为：

$$g = \tanh(P_i : n), g \in \mathbb{R}^{3n} \quad (3)$$

式中 R 为实数域。

主要构成植物的简单化学元素列表包含磷等,与动物列表是一样的。

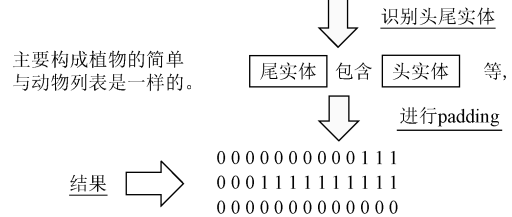


图 7 分段最大池化层描述图

Fig.7 Description of the segmented maximum pooling layer

2.2.3 分类层

为防止过拟合现象以及提升模型的鲁棒性，在分类层处理之前加入 L_2 正则化以及 Dropout 层进一步处理卷积层和池化层的输出。 L_2 正则化与 L_1 正则化的区别是严厉惩罚大数值的权重向量，鼓励使用较小的参数。其中 θ 为即将要学习的参数， λ 控制正则项的大小，在本文中设置为 0.0001。 L_2 正则化对应的损失函数公式为：

$$L = L(\theta) + \lambda \sum_i \theta_i^2 \quad (4)$$

利用线性计算将池化层得到的向量维度降低至 z 维，通过 softmax 分类器预测条件概率并选出最有可能的关系。 t 为关系分类的总类别， \hat{r} 为最终输出结果， softmax 计算公式为：

$$P_i = \frac{e^i}{\sum_{k=1}^t e^k} \quad (5)$$

$$\hat{r} = \arg \max_i (P_i) \quad (6)$$

本研究为多分类问题，采用 Adam 算法作为深度学习中的优化器算法。 Adam 算法可以通过损失函数自动计算一阶矩和二阶矩，从而更新每个参数的学习率。计算公式为：

$$\text{Loss} = -\sum_{k=1}^s T_{kq} \lg p_{kq} \quad (7)$$

式中 s 为关系的总数量， k 为样本种类， q 为分类的类别， p_{kq} 是样本 k 采用 softmax 分类器预测属于 q 的概率值，当预测正确时 T_{kq} 为 1，预测错误时为 0。

2.3 模型改进

2.3.1 注意力机制

Treisman 和 Gelade 提出了一种模拟人脑注意力机制的模型^[34]，该机制可以更好的解决非结构化文本的上下文相关问题，使得训练结果更加准确。注意力机制本质是一种特殊的加权计算，能够过滤出非重要的信息并集中注意力在重要信息上。本文在分段最大池化层与分类层之间添加注意力机制，即在分段池化层之后做进一步的高层语义提取。中间参数 M 的计算方法是将分段最大池化层的输出 P' 压缩到 (-1,1) 值域，权重 W 的计算方法是将矩阵与注意力机制初始化矩阵相乘，最终传到 softmax 函数中得到分类结果 a ，向量 vec 表示加权后的输出向量。其计算公式如下：

$$M = \tanh(P') \quad (8)$$

$$a = \text{softmax}(W^T M) \quad (9)$$

$$\text{vec} = P' a^T \quad (10)$$

2.3.2 基于词粒度的随机掩码

PCNN 模型的分段最大池化层会执行掩码语言模型 (Masked Language Model, MLM)，即在网络训练过程中随机遮盖 (mask) 部分单词，通过上下文输入到 PCNN 网络并进行预测。由于 MLM 过程最开始是针对英文单词的随机掩码方法，若针对每个按照空格区分的单词进行掩码，则不会损失句子含义。文献[35]针对 BERT 模型提出不以字为粒度进行随机掩码切分，而是以分词之后的单词为粒度进行随机掩码，从而提升模型在中文语料库中的准确度。针对食品安全领域中的关系抽取问题，本文利用 Jieba 分词技术对句子分词，分词过程如图 8 所示。对比图 7 和图 8 可以发现，经过分词之后的结果比未经过分词的结果长度更短，如把句子中的“植物”换成较长的植物名称如“常绿木质藤”，若该词被遮盖，不会影响句子结构，且能更好的进行自监督训练。

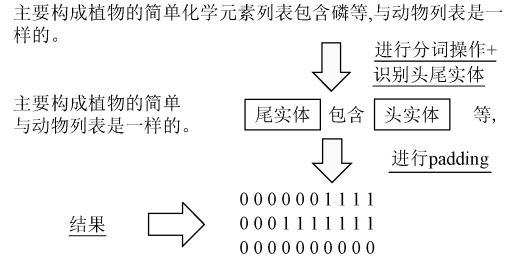


图 8 利用 Jieba 分词的分段池化标注过程图

Fig.8 Process of segmentation pooling tagging using Jieba word segmentation

3 试验与结果

3.1 评价指标

本研究属多分类预测，其评价指标为准确率 (Precision, P , %)，召回率 (Recall, R , %)。每次的训练结果 TP、FP、FN 和 TN，即为正样本被预测正确的数目、正样本被预测错误的数目、负样本被预测错误的数目、负样本被预测正确的数目。准确率和召回率计算公式如 (11)、(12) 所示。当评价不同模型时，准确率和召回率可能各有高低，而无法直接对比多个模型的优劣。因此通过比较 F_1 值 (F_1 value, %) 综合评价各模型，其计算公式如 (13) 所示：

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$F_1 \text{ 值} = \frac{2PR}{P + R} \quad (13)$$

3.2 试验环境与参数设置

试验环境设置为：操作系统为 windows10，采用的处理器是 Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz，内存大小为 12 GB；采用的 GPU 型号是 RTX 3050，显存大小是 4 GB；试验所用的 python 版本为 3.7。

为了充分对比本文提出的 BERT-PCNN-ATT-Jieba 模型，在同一食品安全领域数据集下，将数据集按照 8 : 2 比例划分成训练集和测试集输入模型进行试验。同时选用经典的 CNN、PCNN 模型，结合 BERT 的 PCNN^[19]、PCNN-ATT^[36]和 PCNN-Jieba^[37]模型，以及 BERT-CNN^[20]等 6 个模型作为基准模型进行对比试验。

为控制试验变量，7 个模型的试验参数如表 5 所示。

表 5 神经网络参数设置

Table 5 Parameter settings of neural network

参数 Parameter	数值 Value
学习率	0.000 6
批次大小	64
Dropout 率	0.1
词嵌入向量维度	60
位置嵌入向量维度	40

3.3 试验结果分析

在上述的试验参数配置下，7 个模型对 7 种关系进行

抽取, 不同模型的准确率、召回率、 F_1 值如表 6 所示。

表 6 各神经网络模型准确率、召回率、 F_1 值

Table 6 Precision, recall and F_1 value of each neural network model

模型 Model	准确率 Precision/%	召回率 Recall/%	F_1 值 F_1 value/%
CNN	70.55	71.72	71.13
PCNN	74.93	73.48	74.20
BERT-CNN	76.69	75.13	75.90
BERT-PCNN	78.48	76.96	77.71
BERT-PCNN-ATT	80.87	78.81	79.83
BERT-PCNN-Jieba	83.08	79.95	81.48
BERT-PCNN-ATT-Jieba	84.72	81.78	83.22

从表 6 可以看出, PCNN 模型的 F_1 值比 CNN 模型高, 说明 PCNN 模型更适合于所构建的数据集, 分段最大池化层相较于最大池化层能够捕捉头尾实体的位置关系, 获取丰富的上下文信息, 可以更好的发挥模型优势。BERT-PCNN 与 PCNN 模型相比, BERT-PCNN 的准确率、召回率和 F_1 值均略有提升, 说明利用 BERT 模型生成的向量能够更好的获取数据的语义特征信息。对比 BERT-PCNN-ATT 和 BERT-PCNN 模型发现, 在池化层与分类层之间添加注意力机制后, 可以将池化后的高层语义特征赋予更高的权重值, 说明注意力机制能够提升模型效果。BERT-PCNN-Jieba 模型比 BERT-PCNN 模型的 F_1 值高, 因为在面向食品安全领域的训练集中, 通过对句子的预处理可以减弱单词长度对结果的影响。通过加入分词操作, 能更好的分析词与词之间的位置信息和逻辑信息。与其他模型相比, BERT-PCNN-ATT-Jieba 的准确率、召回率和 F_1 值均为最高, 说明在基于食品安全领域的关系抽取数据集中, 本文提出的 BERT-PCNN-ATT-Jieba 模型相较于其他模型取得更优的性能, 其准确率达到 84.72%, 召回率达到 81.78%, F_1 值达到 83.22%。

BERT-PCNN-ATT-Jieba 模型对抽取不同关系的准确率、召回率及 F_1 值如表 7 所示。

表 7 BERT-PCNN-ATT-Jieba 模型对不同关系的处理结果

Table 7 Processing results of different relationships based on BERT-PCNN-ATT-Jieba model

关系 Relation	准确率 Precision/%	召回率 Recall/%	F_1 值 F_1 value/%
包含	90.43	89.04	89.73
属于	78.06	79.03	78.54
部分	82.61	87.69	85.07
导致	81.98	93.06	87.17
损害	86.73	61.28	71.82
症状	84.11	87.47	85.76
易感人群	85.82	71.97	78.29

从表 7 的结果可以看出, BERT-PCNN-ATT-Jieba 模型在 7 种关系的 F_1 值均高于 70%, 但是对于不同关系的抽取效果有所不同。其中, 对于包含、部分、导致、症状这 4 种关系抽取效果较好, F_1 值高于 85%。而对于属于、损害、易感人群这 3 种关系抽取效果较差, F_1 值小于 80%。

其中包含关系的抽取结果最佳, 其 F_1 值是关系抽取结果最差的损害关系的 1.25 倍。部分关系抽取结果较差的原因是数据量较少以及语料库中英文夹杂现象过多。

4 结 论

1) 本文主要对食品安全领域的关系抽取进行研究, 主要包括构建食品安全领域语料库, 并提出针对该领域的 BERT-PCNN-ATT-Jieba 模型。该模型在 BERT 生成输入词向量的基础上, 利用 PCNN 分段最大池化层能极大程度捕获句子局部信息的优点, 同时结合中文语料的特性, 加入注意力机制与分词处理。

2) 试验结果表明, 在试验参数及数据集一致的条件下, 本文提出的 BERT-PCNN-ATT-Jieba 模型能够更好的抽取基于食品安全领域的语义关系特征, 准确率达到 84.72%, 召回率达到 81.78%, F_1 值达到 83.22%。为下一步构建更完整的食品安全领域知识图谱, 以及基于该知识图谱进行领域知识问答、领域知识检索以及食品安全问题处理等工作提供参考。

3) 本研究的不足之处有, 数据集中语料数量有限且有中英文夹杂现象、标注的关系种类不多、可能会出现分词错误以及未构建重叠三元组提取模型等局限导致部分关系抽取效果受影响。为了达到更好的关系抽取效果, 今后的工作主要集中在数据集的清洗、整理和扩充以及模型的不断改进等方面。

[参 考 文 献]

- [1] Amit S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.
- [2] Paulheim H, Cimiano P. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic Web, 2016, 8(3): 489-508.
- [3] 侯梦薇, 卫荣, 陆亮, 等. 知识图谱研究综述及其在医疗领域的应用[J]. 计算机研究与发展, 2018, 55(12): 2587-2599.
Hou Mengwei, Wei Rong, Lu Liang, et al. Overview of knowledge mapping and its application in the medical field[J]. Computer Research and Development, 2018, 55(12): 2587-2599. (in Chinese with English abstract)
- [4] 夏恩君, 宋剑锋. 开放式创新研究的演化路径和热点领域分析: 基于科学知识图谱视角[J]. 科研管理, 2015, 36(7): 28-37.
Xia Enjun, Song Jianfeng. An analysis of the evolution path and hot topics of open innovation based on the view of the scientific knowledge map[J]. Science Research Management, 2015, 36(7): 28-37. (in Chinese with English abstract)
- [5] 刘焯宸, 李华昱. 领域知识图谱研究综述[J]. 计算机系统应用, 2020, 29(6): 1-12.
Liu Yechen, Li Huayu. Survey on domain knowledge graph research[J]. Computer Systems and Applications, 2020, 29(6): 1-12. (in Chinese with English abstract)
- [6] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.
Liu Jiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600. (in Chinese with English abstract)
- [7] 李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述[J].

- 计算机研究与发展, 2020, 57(7): 1424-1448.
- Li Dongmei, Zhang Yang, Li Dongyuan, et al. Review of entity relation extraction methods[J]. Computer Research and Development, 2020, 57(7): 1424-1448. (in Chinese with English abstract)
- [8] 李涓子, 侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版), 2017, 40(3): 454-459.
- Li Juanzi, Hou Lei. Reviews on knowledge graph research[J]. Journal of Shanxi University (Natural Science Edition), 2017, 40(3): 454-459. (in Chinese with English abstract)
- [9] 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述[J]. 计算机系统应用, 2019, 28(6): 1-12.
- Huang Hengqi, Yu Juan, Liao Xiao, et al. Review on knowledge graphs[J]. Computer System Application, 2019, 28(6): 1-12. (in Chinese with English abstract)
- [10] Chinchor N, Marsh E. Muc-7 information extraction task definition[C]//Proc of the 7th Message Understanding Conf, Philadelphia, USA, 1998: 359-367.
- [11] Aitken J S. Learning information extraction rules: An inductive logic programming approach[C]//Proc of External Credit Assessment Institution, Lyon, France, 2002: 355-359.
- [12] Zhou G, Su J, Zhang J, et al. Exploring various knowledge in relation extraction[C]//The 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Michigan, USA, 2005: 427-434.
- [13] Zelenco D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(2): 1083-1106.
- [14] Jiang J, Zhai C X. A systematic exploration of the feature space for relation extraction[C]//Proc of the Conf of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, USA, 2007: 113-120.
- [15] 王东波, 吴毅, 叶文豪, 等. 多特征知识下的食品安全事件实体抽取研究[J]. 数据分析与知识发现, 2017, 1(3): 54-61.
- Wang Dongbo, Wu Yi, Ye Wenhao, et al. Research on food safety event entity extraction based on multi-feature knowledge[J]. Data Analysis and Knowledge Discovery, 2017, 1(3): 54-61. (in Chinese with English abstract)
- [16] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.
- E Haihong, Zhang Wenjing, Xiao Siqi, et al. Survey of entity-relationship extraction based on deep learning[J]. Journal of Software, 2019, 30(6): 1793-1818. (in Chinese with English abstract)
- [17] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]//The 25th International Conference on Computational Linguistics, Dublin, Ireland, 2014: 2335-2344.
- [18] 王庆棒, 汪颖懿, 左敏, 等. 基于 CNN-BLSTM 的食品舆情实体关系抽取模型研究[J]. 食品科学技术学报, 2021, 39(2): 152-158.
- Wang Qingbang, Wang Haoyi, Zuo Min, et al. Research on entity-relationship extraction model of food public opinion based on CNN-BLSTM[J]. Journal of Food Science and Technology, 2021, 39(2): 152-158. (in Chinese with English abstract)
- [19] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015: 1753-1762.
- [20] 武小平, 张强, 赵芳, 等. 基于 BERT 的心血管医疗指南实体关系抽取方法[J]. 计算机应用, 2021, 41(1): 145-149.
- Wu Xiaoping, Zhang Qiang, Zhao Fang, et al. Entity relation extraction method for guidelines of cardiovascular disease based on bidirectional encoder representation from transformers[J]. Journal of Computer Application, 2021, 41(1): 145-149. (in Chinese with English abstract)
- [21] Wang J, Guo Y. Scrapy-based crawling and user-behavior characteristics analysis on taobao[C]// 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Sanya, China, 2012: 44-52.
- [22] Percuku A, Minkovska D, Stoyanova L. Modeling and processing big data of power transmission grid substation using Neo4j[J]. Procedia Computer Science, 2017, 113: 9-16.
- [23] Holzschuher F, Peinl R. Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4j[C]//Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, Italy, 2013: 195-204.
- [24] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: One hot way to resist adversarial examples[C]//International Conference on Learning Representations, British Columbia, Canada, 2018: 1-22.
- [25] Mikolov T, Corrado G, Kai C, et al. Efficient Estimation of Word Representations in Vector Space[C]//International Conference on Learning Representations, Arizona, USA, 2013: 1-12.
- [26] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of NAACL-HLT, Louisiana, USA, 2018: 2227-2237.
- [27] 李舟军, 范宇, 吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020, 47(3): 170-181.
- Li Zhoujun, Fan Yu, Wu Xianjun. A review of research on pre-training technology for natural language processing[J]. Computer Science, 2020, 47(3): 170-181. (in Chinese with English abstract)
- [28] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv, 2018, 1810.04805.
- [29] 陈德光, 马金林, 马自萍, 等. 自然语言处理预训练综述[J]. 计算机科学与探索, 2021, 15(8): 1359-1389.
- Chen Deguang, Ma Jinlin, Ma Ziping, et al. Review of pre-training techniques for natural language processing[J]. Computer Science and Exploration, 2021, 15(8): 1359-1389. (in Chinese with English abstract)
- [30] 郑丽敏, 任乐乐. 采用融合规则与 BERT-FLAT 模型对营养健康领域命名实体识别[J]. 农业工程学报, 2021, 37(20): 211-218.
- Zheng Limin, Ren Lele. Named entity recognition in human nutrition and health domain using rule and BERT-FLAT[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(20): 211-218. (in Chinese with English abstract)
- [31] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于 BERT 的多特征融合农业命名实体识别[J]. 农业工程学报, 2022, 38(3): 112-118.
- Zhao Pengfei, Zhao Chunjiang, Wu Huarui, et al. Recognition of the agricultural named entities with multi-feature fusion based on BERT[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(3): 112-118. (in Chinese with English abstract)
- [32] 任媛, 于红, 杨鹤, 等. 融合注意力机制与 BERT+BiLSTM+CRF 模型的渔业标准定量指标识别[J]. 农业工

- 程学报, 2021, 37(10): 135-141.
- Ren Yuan, Yu Hong, Yang He, et al. Recognition of quantitative indicator of fishery standard using attention mechanism and the BERT+BiLSTM+CRF model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(10): 135-141. (in Chinese with English abstract)
- [33] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv, 2014,1412,06980.
- [34] Mnih, Volodymyr, Heess, et al. Recurrent models of visual attention[J]. arXiv, 2014,1406,06247.
- [35] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J]. arXiv, 2019,1906. 08101.
- [36] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 2124-2133.
- [37] 张彤, 宋明艳, 王俊, 等. 基于 PCNN 的工业制造领域质量文本实体关系抽取方法[J]. 信息技术与网络安全, 2021, 40(3): 8-13.
- Zhang Tong, Song Mingyan, Wang Jun, et al. Entity relation extraction method for quality text in industrial manufacturing field based on PCNN[J]. Information Technology and Network Security, 2021, 40(3): 8-13. (in Chinese with English abstract)

Relationship extraction in the field of food safety based on BERT and improved PCNN model

Zhao Liang^{1,2}, Zhang Zhaoyue³, Liao Ziyi⁴, Wang Ling^{1,2}

(1. College of Informatics, Huazhong Agricultural University, Wuhan 430070, China; 2. Hubei Engineering Technology Research Center of Agricultural Big Data (Huazhong Agricultural University), Wuhan 430070, China; 3. School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China; 4. Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: A knowledge graph (semantic network) has emerged to organize the real-world entities in a graph database for the relationship between them. Among them, relationship extraction has been one of the most important links in the automatic construction of knowledge graphs. However, there is no public dataset related to knowledge graphs in the food safety field at present. The existing models of relationship extraction are confined to the open standard data set, but most cannot extract the data in the specific domain. In this study, a professional data set was constructed for the relationship extraction in the food safety field using the Bidirectional Encoder Representations from Transformers (BERT) and the improved Piecewise Convolutional Neural Network (PCNN) model. The corpus was firstly collected to annotate the corresponding entities and related categories. At the same time, a relationship extraction model was proposed using BERT-PCNN-Attention-based Neural Networks (ATT)-Jieba for the field of food safety. The BERT pre-training model was selected to generate the input word vector. After that, the segmented maximum pooling layer of the PCNN model was utilized to capture the local information of sentences. An attention mechanism was added between the segmented maximum pooling layer and the classification layer, further to extract the high-level semantics. In addition, Jieba word segmentation was used to segment the Chinese corpus before the random mask segmentation of the BERT model. The segmented maximum pool layer of the PCNN model masked the word unit instead of characters when executing the Masked Language Model (MLM). As such, the semantic loss of sentences was reduced to achieve a more efficient relationship extraction, when inputting into the training model. The performance of the BERT-PCNN-ATT-Jieba model was compared with the classical CNN, PCNN model, as well as the CNN, PCNN, PCNN-ATT, and PCNN-Jieba models combined with BERT under the same data set and the consistent experimental parameters. Comparing the PCNN with the BERT-PCNN model, the precision, recall, and F_1 value of BERT-PCNN were slightly improved, indicating that the vector generated by the BERT model can better obtain the semantic feature information of data. Comparing the BERT-PCNN-ATT and BERT-PCNN, the pooled high-level semantic features presented a higher weight value after adding the attention mechanism between the pooling layer and the classification layer, indicating that the attention mechanism can improve the performance of the model. The F_1 value of BERT-PCNN-Jieba was better than that of BERT-PCNN because the influence of word length was weakened through sentence preprocessing in the training set for the field of food safety. The position and logical information between words were better analyzed by adding a word segmentation operation. Consequently, the BERT-PCNN-ATT-Jieba model presented the highest precision of 84.72%, recall of 81.78%, and F_1 value of 83.22%, indicating that the better performance was achieved in the relationship extraction data set using the field of food safety. The finding can provide a strong reference for knowledge extraction in the cost-saving and automatic construction of knowledge graphs in the field of food safety. The improved model can also lay a foundation for the application of Knowledge Q&A, knowledge retrieval, data sharing, and intelligent supervision of food safety using knowledge graphs.

Keywords: food safety; models; relationship extraction; knowledge graph; attention mechanism; BERT; PCNN