文章编号：2095-2163(2022)02-0032-06 中图分类号：TP391 文献标志码：A

面向新兴产业的检验检测服务关系抽取

张婷婷[1]，让 冉[1]，张龙波[1]，邢林林[1]，蔡红珍[2]
(1 山东理工大学 计算机科学与技术学院，山东 淄博 255000)；2 山东理工大学 农业工程与食品科学学院，山东 淄博 255000)

摘 要：挖掘新兴产业中的检测信息有利于发现检测机构的检测能力，促进机构合作，加速产业升级。针对新兴产业检验检测数据存在语义混乱、一个主实体对应多个客实体的问题，本文提出一种混合关系标签的深度神经网络实体关系抽取模型。在输入层，构建基于关系的标签，并与语义信息拼接形成模型的输入，增强了模型对不同关系的区分度；在特征提取层，使用双向长短期神经网络与卷积神经网络，从整体与局部提升模型对主客实体特征的抽取能力，同时引入注意力机制，削弱无关特征的影响，提升模型对主客实体的识别能力。实验结果表明，该模型不仅能有效识别出新兴产业检验检测领域的实体，而且能精准判断实体之间的关系，取得了较好的结果。

关键词：新兴产业；检验检测；关系抽取；标签；卷积神经网络

Detection service relation extraction for emerging industries

ZHANG Tingting[1], RANG Ran[1], ZHANG Longbo[1], XING Linlin[1], CAI Hongzhen[2]
(1 College of Computer Science and Technology, Shandong University of Science and Technology, Zibo Shandong, 255000, China;
2 College of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo Shandong, 255000, China)

【Abstract】 Mining detection information in emerging industries is conducive to discovering detection capabilities of detection institutions, promoting cooperation between institutions and accelerating industrial upgrading. In order to solve the problem of semantic confusion in the inspection data of emerging industries, a deep neural network entity relation extraction model based on mixed relation labels is proposed in this paper. At the input layer, the labels based on relationships are constructed and the input of the model is spliced together with semantic information to enhance the differentiation of different relationships. At the feature extraction layer, bidirectional long and short-term neural network and convolutional neural network are used to improve the feature extraction ability of the model. Meanwhile, attention mechanism is introduced to weaken the influence of irrelevant features and improve the recognition ability of the model to host and object entities. The experimental results show that the model can not only identify the entities in the inspection and testing field of emerging industries, but also accurately judge the relationship between entities, achieving good experimental results.

【Key words】 emerging industries; inspection and testing; relation extraction; labels; convolution neural network

0 引 言

新兴产业是随着新的科研成果和新型技术的诞生，应运而生的新的经济部门或行业，涉及节能环保、生物产业、新能源、新材料等众多领域。检验检测是指通过专业技术手段对动植物、工业产品、商品、专项技术、成果及其他需要鉴定的物品所进行的检测、检验、测试、鉴定等活动。对新兴产业检验检测数据研究发现其中包含了大量有用的信息，包括各检测机构可提供的检测范围及机构地理位置信息等，将这些信息结构化保存，可以为领域内的合作关系提供数据指导，有利于后续的产业升级。关系抽取是自然语言处理(Natural Language Processing, NLP)的基层任务之一，主要是从文本中识别出实体，并抽取实体之间的语义关系[1]。在新兴产业领域使用关系抽取技术能够快速发现机构可提供的检测项目，继而构建新兴产业检验检测关系集，不仅能够为寻求检测服务的机构提供精准的业务推荐，还能总览行业整体发展水平，促进新兴产业快速发展。

新兴产业检验检测关系抽取缺少专业的实验数据，且数据中广泛存在语义混乱、一个主实体对应多个客实体的问题，基于此本文构建了领域内数据集并提出了一种适用于新兴产业检验检测数据的关系抽取模型，主要工作包括：

基金项目：国家重点研发计划资助项目(2018YFB1403302)。

作者简介：张婷婷(1995-)，女，硕士研究生，主要研究方向：自然语言处理；让 冉(1998-)，女，硕士研究生，主要研究方向：自然语言处理；张龙波(1968-)，男，博士，教授，硕士生导师，主要研究方向：数据挖掘；邢林林(1987-)，男，博士，讲师，主要研究方向：生物信息学；蔡红珍(1972-)，女，博士，教授，博士生导师，主要研究方向：复合材料。

通讯作者：邢林林 Email：xinglinlin@ sdut.edu.cn

收稿日期：2021-10-18

（1）在输入层为每个关系设置不同的关系标签，并与字词向量、位置向量混合作为神经网络架构模型的输入，其中位置向量能从主实体及客实体两个层面分析语义位置信息对关系判定的影响；标签向量可以强化各关系的界限，有利于后续充分学习各关系的特征信息，提高分类准确率；

（2）提出由卷积神经网络（Convolutional Neural Network，CNN）和双向长短期记忆网络（Long Short Term Memory，LSTM）组成的关系抽取模型，既能够利用 BiLSTM 兼顾长序列文本的整体信息，又能够利用 CNN 提取文本局部有价值的特征，同时使用选择性注意力机制有针对性对字词赋予不同的权值大小，提升模型在特征提取方面的准确率。

1　相关工作

传统的关系抽取主要基于模式匹配和基于机器学习的研究，基于模式匹配的方法需要制定关系抽取规则，耗时耗力；基于机器学习的方法利用数据训练算法，但是不能充分提取数据文本中的语义信息[2]。深度学习最重要的特点就是可以反向传播学习，同时自动学习实验数据中的特征。将深度学习应用到关系抽取中是目前计算机领域的研究重点，主要以 CNN、LSTM、循环神经网络（Recurrent Neural Network，RNN）、双向门控循环单元（Gate Recurrent Unit，GRU）等结构来展开[3]。Liu 等[4]构造了一个由输入层、卷积层、池化层及最后由 Softmax 分类器输出模型的分类结果的 CNN 神经网络结构，证明了 CNN 在提取特征方面有良好的效果，并将其应用到自然语言处理领域；Zeng[5]等人将 CNN 应用到关系抽取过程中，同时提取了句子文本中的语义特征；Elman 提出了第一个全连接的 RNN 网络结构后，Socher[6]等将 RNN 模型用于关系抽取领域，取得了阶段性的进展。但是 RNN 在处理长的序列文本时同样也带来了严重的问题：梯度消失、梯度爆炸以及可能存在训练时间过长的问题，由此提出了 LSTM[7]；Zhang 等人[8]采用 BiLSTM 模型结合文本的前后语义进行抽取，提升了模型的准确率。注意力机制的目标是以概率的形式对文本中的特征进行重要性的区分，Mnih 等[9]将 RNN、注意力机制与张量层相结合的关系抽取分类算法，有效提高了分类的结果；Cai 等[10]以 CNN 和 LSTM 为基础，设计使用双向 CNN，同时从两个方向学习最短依存信息，取得了阶段性新进展。

不同于英文领域存在大规模的 ACE2004 实验语料、NYT-FB 数据集等专业语料库，受限于标准化语料库的规模以及数量，中文关系抽取还存有很大的进步空间。在中文领域中，生物医疗的应用尤为广泛，文献[11]提出一种基于双向 GRU 和 CNN 相融合的双层药物关系抽取模型，在医药领域专门数据及 DDIExtraction2013 中取得了 75% 的综合测评率；文献[12]在面对高密度分布的实体以及实体间关系的交叉互联为电子病历时，提出一种基于多通道自注意力机制的 BiLSTM+多通道自注意力机制的神经网络架构，$F1$ 值最高提升了 23%；其他领域，如文献[13]利用高质量的食品安全事件新闻文本完成领域内实体关系抽取工作；文献[14]在社交媒体领域提出了一种基于注意力机制以及 LSTM 的好友关系预测模型，实验结果的准确率达到了 77%；文献[15]针对煤矿领域数据存在的术语嵌套、一词多义等问题，基于 BiLSTM 提出了一种端到端的联合学习模型，提高了模型的训练速度；文献[16]根据老挝语料库匮乏的特点，结合“硬匹配”和“软匹配”的思想，提出了一种基于双向长短期记忆网络和多头自注意力机制的军事领域实体关系抽取方法，都取得了很好的实验效果。

2　模型设计

本文提出了一个面向新兴产业检验检测领域的神经网络架构模型，结构如图 1 所示。

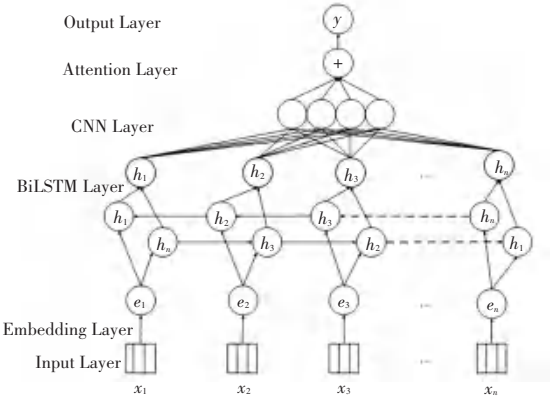

图 1　检验检测神经网络架构模型图

Fig. 1　The architecture of detection neural network model

其核心流程如下：

（1）将数据文本中的字词集合转化为包含字词信息、位置信息、标签信息的向量表示；

（2）将获取到的向量传入由 BiLSTM 和 CNN 组成的深度神经网络模型，深度学习每个关系的语义特征信息；

（3）使用注意力机制削弱无关特征的影响；

(4)将关系抽取转化为分类任务,输出分类结果。

2.1 输入向量

输入向量主要包括 3 部分:词向量,用来将检验检测领域内的字词转化为向量表示;位置向量,根据文本中字词相对于主实体的距离进行标记;标签向量,根据句子中实体间关系的不同类别,设置相应的标签向量。

2.1.1 词向量

将实验所用的领域文本 $S=\{s_1,s_2,\cdots,s_n\}$,通过公式(1)将每个字词 $\boldsymbol{s}_i$ 映射到低维的向量空间中,构建出每个字词特征向量,然后对各个词向量进行拼接,形成此次实验的字词向量,式(1)。

$$\boldsymbol{s}_i=\boldsymbol{w}^{word}\cdot v^i \tag{1}$$

其中,$\boldsymbol{w}^{word}$ 是通过 Word2vec 得到的词向量矩阵,v_i 是词 $\boldsymbol{s}_i$ 的 one-hot 表示。

2.1.2 位置向量

在关系抽取领域,一般越是靠近实体的字词越能准确的反映数据中实体之间的关系。为了更准确的描述文本中主实体的关系,本文构建了位置向量 $\boldsymbol{p}_i$,分别从前后方向两个不同的角度来表示每个字到实体间的距离,式(2)。

$$\boldsymbol{p}_i=[\overrightarrow{p_i},\overleftarrow{p_i}] \tag{2}$$

以“华测检测积累了较完整的机器人检测经验”中“华测检测”为例,向量表示为 $[\overrightarrow{0,1,2,3},\overleftarrow{-11,-10,-9,-8}]$,剩余部分位置向量如图 2 所示。

图 2 位置向量示意图

Fig. 2 Position vector diagram

2.1.3 标签向量化

针对事先定义好的几种分类结果,分别建立相应的标签,并转化为向量的形式;最后,将标签向量、字词向量与位置向量拼接在一起形成本次实验的输入向量,如公式(3)所示。

$$\boldsymbol{x}_i=\{\boldsymbol{s}_i,\boldsymbol{p}_i,l_i\} \tag{3}$$

2.2 神经网络结构

2.2.1 BiLSTM

LSTM 能够在有效利用长距离信息的同时解决 RNN 存在的梯度消失或者梯度爆炸的问题,LSTM 结构如图 3 所示。

LSTM 由遗忘门 f_t、输入门 i_t 和输出门 o_t 以及一个记忆单元组成,其中遗忘门决定什么样的信息应该被神经元遗忘,输入门决定什么样的信息应该被神经元输入,输出门决定什么样的信息应该被神经元输出,记忆单元用来管理和保存神经元中的参数信息,计算过程如式(4)~式(6)所示:

$$i_t=\sigma(W_{xi}\boldsymbol{x}_i+W_{hi}x_{t-1}+b_i) \tag{4}$$

$$f_t=\sigma(W_{xf}\boldsymbol{x}_i+W_{hf}h_{t-1}+b_f) \tag{5}$$

$$o_t=\sigma(W_{xo}\boldsymbol{x}_i+W_{ho}h_{t-1}+b_o) \tag{6}$$

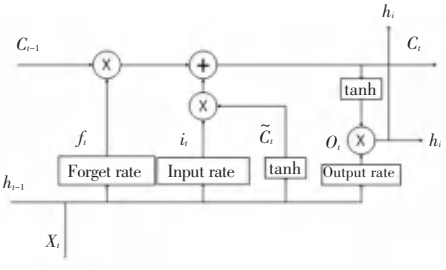

图 3 LSTM 结构图

Fig. 3 LSTM structure diagram

文本中的每个字词都蕴含着重要的信息,使用 BiLSTM 来获取数据中的特征信息。BiLSTM 均与输出层相连,分别作用于文本的前后文信息,在长序列文本数据集上有很好的表现效果。

2.2.2 CNN

CNN 主要由输入层、卷积层、池化层和 Dropout 组成,卷积层通过设置不同规模的卷积核提取检验检测领域内不同的文字特征,且共享卷积核参数,公式(7)表示其特征的提取过程,得到特征集合 $f=(f_1,f_2,f_3,\cdots,f_n)$。

$$f_i=\tanh(t\cdot h_{i:i+k-1}+b) \tag{7}$$

其中,b 表示偏差,tanh 为双曲正切函数。

池化层通过对数据进行压缩来实现数据和参数的降维,从而降低模型过拟合的概率,本次实验采用最大池化方法,取 f_n 中最大的 f_i,挑选最明显的特征。全连接层主要是把分布式特征映射到样本标记空间中,对之前提取到的特征进行分类。Dropout 在训练模型时按照一定的概率将一部分神经网络暂时丢弃,不仅能防止模型过拟合还可以提高网络的泛化能力。

在本次试验中,通过双向 LSTM 神经网络从前后两个方向全面获取数据文本的文本序列信息,然后将向量拼接传入 CNN 中,进一步提取文本中的关键特征。

2.2.3 注意力机制

将预测关系转化为文本分类过程中,不同字词发挥着不同的作用,使用注意力机制可以为数据中的字词设置相应的权重,对分类起正向作用的字词

设置更高的权重，增强正向字词的影响。将 BiLSTM 与 CNN 模型提取后形成的词向量矩阵 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_n\}$ 经过 tanh 函数映射到[-1,1]之间，再将映射结果传入 softmax 函数，计算得到每个词的注意力分数，最后对句子的各个单词的编码结果进行加权求和，其计算过程如式(8)~式(10)所示：

$$M=\tanh(C) \tag{8}$$

$$\alpha=\mathrm{softmax}(W^{\mathrm{T}}M) \tag{9}$$

$$r=H\alpha^{\mathrm{T}} \tag{10}$$

3　实验结果与分析

3.1　实验数据

在实验之前，本文对新兴产业检验检测数据研究发现，领域内的文本实体数量相对复杂，以“无锡检验检测中心的检测业务包括节能环保、生物医药、新材料、新能源等重要领域”为例，全句共包含 5 个实体，4 种检测关系，其中无锡检验检测中心同时对应着 4 个不同的实体，其关系如图 4 所示，提升了关系抽取难度。

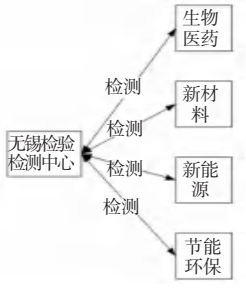

图 4　数据关系示意图

Fig. 4　Data relationship diagram

本次实验共选取 1 800 条新兴检验检测语料，按照 4∶1 的比例划分为训练数据与测试数据，其中训练语料中包含 7 192 个实体，测试语料中包括 1 690个实体，共计 8 882 个实体。通过观察发现，新兴检验检测行业能否达成合作的关键在于各检测机构可以提供的检测服务类型以及检测机构的位置。基于此，本次试验着重抽取“位置”、“检测”两大类的数据信息，关系示例见表 1。

表 1　关系示意表

Tab. 1　Relationship diagram

示例文本	实体	关系
谱尼测试集团可开展汽车轮胎检测业务	谱尼测试集团、汽车轮胎	检测
谱尼测试集团于北京建立分公司	谱尼测试集团、北京	位置

3.2　模型训练与优化

为了使模型作用最大化，本文进行了一系列实验来确定最合适的参数数值，包括词随机丢弃率、LSTM 随机丢弃率、Attention 随机丢弃率与学习率等，根据实验结果，各参数的数值见表 2。

表 2　参数示意表

Tab. 2　Statement of parameters

参数名	参数值
Word Dropout probability	0.3
LSTM Dropout probability	0.5
ATT Dropout probability	0.5
Position dimension	80
Word size	256
Learning_rate	0.000 5

实验采用 Adam 优化器来迭代更新神经网络的参数，通过计算梯度的一阶矩估计和二阶矩估计为不同的参数设计独立的自适应性学习率。以交叉熵作为实验的损失函数，熵是对不确定性的度量，交叉熵损失函数通过设置 0 或者 1 的类别标签来显示分类的类别结果，衡量 y 正确值平均起来的不确定性。如果输出的是期望的结果，不确定性就会小一点，交叉熵就越小，如公式(11)和公式(12)所示：

$$S(A)=-\sum_i P_A(x_i)\log P_A(x_i) \tag{11}$$

$$H(A,B)=-\sum_i P_A(x_i)\log(P_B(x_i)) \tag{12}$$

3.3　实验结果

3.3.1　实验结果

本文以平均准确率(P)、平均召回率(R)以及 $F1$ 值的大小来作为评价模型的标准。在上述各参数的配置下，在迭代 100 次以后，本次实验的准确率、召回率以及 $F1$ 值的结果如图 5 所示，可以看出，模型迭代次数从 70 次开始逐渐收敛，模型各指标趋于稳定。

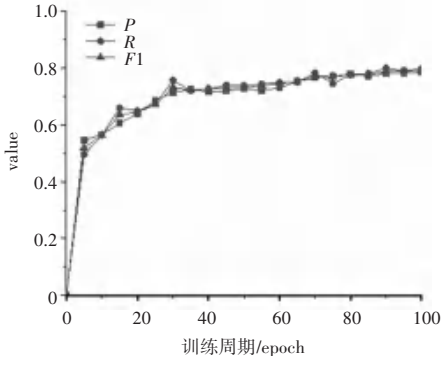

图 5　实验结果

Fig. 5　Experimental results

3.3.2 输入实验对比

为了验证本次实验所提出的四层输入的有效性,在本次实验模型及相同的关系抽取模型的基础上,与普通的输入进行了实验结果对比,见表 3。可以发现本文提出的四层输入相较于普通的输入,准确率、召回率以及 $F1$ 值均有所提升,其中召回率与 $F1$ 值都有 1.9 个百分点的提升。

表 3 结果对比

Tab. 3 Comparison of results %

类型	P	R	$F1$
四层输入	78.3	79.8	79.1
普通输入	76.6	77.9	77.2

3.3.3 实验模型的选择

分别以 BiLSTM + Attention 与 LSTM + CNN + Attention 关系抽取为实验对比模型,在新兴检验检测领域的数据上展开关系抽取实验,结果见表 4。

表 4 结果对比

Tab. 4 Comparison of results %

类型	P	R	$F1$
BiLSTM+CNN+ATT	78.3	79.8	79.1
LSTM+CNN+ATT	73.2	74.3	74.5
BiLSTM+ATT	71.5	74.8	73.7

由此,可以得出以下几个结论:

(1)在同样的实验背景下,双向 LSTM 能够兼顾前后两个不同方向的信息,实验准确率高于单向的 LSTM;

(2)在同样的实验背景下,CNN 利用不同规模的卷积核分别提取实验数据里的特征信息,在 BiLSTM+ATT 模型的基础上加入 CNN 后实验结果有小范围提升;

(3)结合新兴检验检测数据具有文本序列长、语义混乱的特点,BiLSTM 能够兼顾长文本中的序列信息,但由于实体密集,添加引入 CNN 之后,通过设置卷积层能够提取数据中丰富的特征信息,最后利用注意力机制为各特征赋予合适的权重,提升了模型整体分类的准确率。

3.3.4 在人物关系抽取数据集上的验证

为了验证本文所提出的实体关系抽取模型在其他领域的表现效果,选择人物抽取数据集作为实验对照组。将百度公开人物关系抽取数据集经过处理转化为本模型所需要的数据形式,共选取了十大类 50 000条数据进行实验,其中 $F1$ 的损失率如图 6 所示。在迭代 100 次以后,结果逐渐趋于稳定,损失最低为 21.4%,即 $F1$ 值最高达 78.6%,与本文取得的实验结果差别不大。相较于人物关系数据,无论是从文本的序列长度还是相同句子长度下所含有的实体数量,新兴检验检测数据都更为复杂,而这一点恰恰验证了本文提出的模型在相同单位体量的数据集下处理复杂语句的能力。

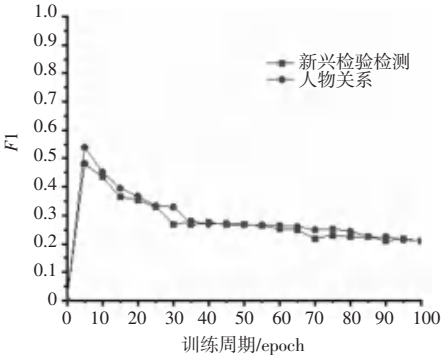

图 6 实验对比结果

Fig. 6 Comparison of experimental results

4 结束语

在新兴检验检测领域内展开关系抽取可以发现机构可提供的检测服务类型以及检测机构的位置信息,能够为后续机构之间的合作提供数据支持,加速机构之间的合作化进程。根据新兴产业检验检测领域数据的特点,本文提出了一种将 BiLSTM 与 CNN 结合起来的领域关系抽取模型。实验结果证明 LSTM 在长序列文本中有较好的表现效果,对于特征混乱且多的句子,加入 CNN 能显著提升特征提取的效果。

未来,随着深度学习的发展,NLP 领域内也会出现更多更简洁、高效的关系抽取算法模型,也可以从模型自身及其他方面进行下一步有针对性的研究,如深入利用字词之间的依存关系或在模型中加入最短依赖或进行迁移学习等等。

参考文献

[1] 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.

[2] 尹鹏,周林,郭强,等. 基于短语级注意力机制的关系抽取方法[J]. 计算机技术与发展,2019,29(9):24-30.

[3] 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.

[4] LIU C Y, SUN W B, CHAO W H, et al. Convolution neural network for relation extraction[C]//International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2013: 231-242.

[5] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 2335-2344.

(下转第 43 页)