第 46 卷 第 5 期 计 算 机 工 程 2020 年 5 月
Vol. 46 No. 5 Computer Engineering May 2020

·开发研究与工程应用· 文章编号：1000-3428(2020)05-0312-09 文献标志码：A 中图分类号：TP391.4

面向财税领域的实体识别与标注研究

仇 瑜[1,2,3]，程 力[1,2,3]

（1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；2. 中国科学院大学，北京 100049；
3. 新疆民族语音语言信息处理实验室，乌鲁木齐 830011）

摘 要： 特定领域中的实体结构和类别相比通用领域更加复杂多样，传统的命名实体识别方法难以取得理想效果。针对该问题，以财税领域为例，研究领域实体识别与标注问题，实现知识库的动态扩充。根据领域特征定义一组层次实体类别集，使用远程监督的方法获取训练语料。采用基于字、词特征结合的深度神经网络模型识别实体边界，将实体类别标注视为多标签多类别分类任务，并提出一种基于集成学习的方法以进行实体类别标注。在真实数据集上的实验结果表明，相比逻辑回归、支持向量机等方法，该方法的准确率、召回率及 F 值更高。

关键词： 知识库扩充；实体识别；实体标注；深度学习；集成学习

开放科学（资源服务）标志码（OSID）：

中文引用格式： 仇瑜，程力. 面向财税领域的实体识别与标注研究[J]. 计算机工程，2020，46(5)：312-320.

英文引用格式： QIU Yu，CHENG Li. Research on entity recognition and tagging in fiscal and taxation domain[J]. Computer Engineering，2020，46(5)：312-320.

Research on Entity Recognition and Tagging in Fiscal and Taxation Domain

QIU Yu[1,2,3]，CHENG Li[1,2,3]

（1. The Xinjiang Technical Institute of Physics and Chemistry，Chinese Academic of Sciences，Urumqi 830011，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China；
3. Xinjiang Laboratory of Minority Speech and Language Information Processing，Urumqi 830011，China）

【Abstract】 Traditional recognition methods for named entities do not work well for entities in specific domains，as they usually have more complex structures and types than those in the general domain. To address the problem，this paper takes the fiscal and taxation domain as an entry point to study entity recognition and tagging，so as to implement dynamic expansion of knowledge base. According to the characteristics of the fiscal and taxation domain，a hierarchical entity type set is defined，and a training corpus is obtained by using remote monitoring. Then a deep neural network model based on combined character features and word features is used for entity boundary recognition. Entity type tagging is taken as a multi-label and multi-type classification task，and on this basis a method based on ensemble learning is proposed for entity type tagging. Experimental results on real datasets show that compared with basic methods including logistic regression and support vector machine，the proposed method has higher accuracy，recall and F value.

【Key words】 knowledge base expansion；entity recognition；entity tagging；deep learning；ensemble learning

DOI： 10.19678/j.issn.1000-3428.0054483

0 概述

随着对谷歌知识图谱研究和应用的深入，知识库成为理解自然语言中实体和关系的背景信息。基于知识库的各类智能应用（如检索、推荐、问答等）也得到了广泛研究，并已取得显著效果[1-2]。目前通用知识库如 YAGO、DBpediat、NELL 和 FreeBase 等已经形成较大的规模，但是这些知识库仍然不够完备，需要不断对其缺失信息进行补充。

近年来，为解决信息缺失问题，知识库扩充技术得到广泛关注。在通用领域，相关研究主要从非结构化的文本中抽取三元组信息（实体，关系，实体），如 DeepDive、NELL 等[3-4]。这些研究主要关注实体和实体之间的关系抽取，对实体类别标注考虑较少

基金项目： 国家“千人计划”项目（Y32H251201）；中国科学院“西部之光”基金（2017-XBZG-BR-001）。
作者简介： 仇 瑜（1988—），男，博士研究生，主研方向为人工智能、自然语言处理；程 力，研究员、博士生导师。
收稿日期： 2019-04-03 **修回日期：** 2019-05-15 **E-mail：** qiuyu12@mails.ucas.ac.cn

(或仅涉及较为有限的实体类别)。通用领域知识库较重视知识的广度,而特定领域知识库强调知识的深度,其更加关注实体的类别信息。相较于通用领域,特定领域的实体类别更加丰富,而且具有较深的层次结构。

传统的命名实体是指现实世界中某个对象的指称(如人名、地名、机构名等),在近期的一些研究中,命名实体的范围更加广泛,实体类别粒度也更细致。如在某些特定领域(如金融、医疗、生物等),命名实体还可以指商品名称、会议名称、疾病名称等[5]。更广泛而言,实体还包括某类实物的总称,如"大象"是一类动物的总称,而不是特定的某个动物。在财税领域中,命名实体涉及更加复杂的实例,如"征税对象"是一个范畴较高的实体,在实际应用中需要识别某类实体而不是具体的某个实体,如"建筑物/办公楼"可以当成一个实体,而不是某个具体地址的楼房。特定领域中的实体类别较多,而且分类更细,传统的命名实体识别方法无法对特定领域实体进行准确地识别与标注。

实体类别预测可以看作分类任务,其目的是将已识别出的候选实体分类到预定义的实体类别中。一般采用有监督学习的方法从标注语料中抽取实体特征,训练分类器然后进行实体类别预测,但是这种方法仍然存在如训练语料不平衡和机器学习算法偏见等问题[6]。此外,相比于通用领域,特定领域语料(如财税法规和案例)中的实体识别和标注还面临 2 个方面的挑战,一为特定领域缺少相关的标注语料和资源,二为目前的实体识别工具主要适用于通用领域,应用于特定领域时效果不理想。

本文构建细粒度的领域类别集和训练语料,并研究适用于特定领域的实体识别方法和细粒度实体标注方法。

1　相关工作

文献[7]提出一种基于卷积神经网络联合模型的细粒度实体标注方法,以对知识库进行扩充。该方法取得了较好的效果,但它是面向通用领域的实体标注,标注模型无法利用领域实体的特征,而且文献中没有关于实体边界识别方法的描述。

在实体边界识别研究中,早期学者多使用词性标注和句法依存关系将名词短语作为候选实体[8],这种方法只能识别形式比较规范的实体。目前,比较通用的方法是将实体识别看成序列标注任务,构建标注器(如条件随机场模型),通过特征模板训练实体特征,以对实体边界进行标注[9-10]。近年来,研究人员开始使用深度学习模型进行序列标注,并已取得较好的效果[11-13]。但是这些方法主要应用于通用领域,针对特定领域实体识别的相关研究还相对较少。

实体标注首先需要定义一组实体类别集,文献[10]从知识库 Freebase 中定义了一组包含 112 个类别标签的扁平化实体类别集 FIGER。文献[8]进一步将 FIGER 的类别集组织成一种层次结构以对实体进行类别标注。文献[14]提出一个包含 505 个类别标签的 5 层深度的层次化类别集,该类别集的定义借鉴知识库 YAGO、Wikipedia 和 WordNet 的结构。目前,在财税领域几乎没有可用的相关资源,需要研究领域知识特征并定义领域实体标注集。

实体标注方法多数采用分类算法,文献[10]使用了线性分类器对实体类别进行预测,特征集包括词法、上下文、聚类等特征。文献[14]利用 SVM 分类器对实体进行分类,使用字符、语法、词典、上下文等特征。文献[8]使用局部分类器和扁平分类器进行分类,并对比测试了两者的分类效果,特征集选用实体的词法、句法及文档主题特征。由于领域实体的特殊性和复杂性,单一的分类器通常泛化能力较差,难以取得令人满意的效果。

本文提出一种针对特定领域的实体识别与标注方法。定义财税领域实体标注集,通过远程监督的方法构建针对实体识别和细粒度实体标注任务的语料集。对传统基于词向量的方法进行改进,提出一种字符特征与词向量相结合的方法并构建神经网络模型,提高领域实体的识别效果。根据领域实体类型多样、结构复杂的特点,提出一种基于集成学习的层次分类算法,以进行实体类别标注。

2　研究方法

2.1　问题定义

本文主要关注中文财税领域的实体识别与标注问题。定义一组领域实体类别集,从非结构化的领域文本中识别出领域实体,并对实体所属的类别进行预测,最后根据标注结果将识别出的实体加入到现有知识库中。上述过程形式化表示为:输入领域文档集 $D=\{d_1,d_2,\cdots,d_i,\cdots,d_n\}$、领域知识库 K_d 以及从知识库中定义的一组类别集 $T=\{t_1,t_2,\cdots,t_i,\cdots,t_n\}$,需要从领域文本中识别出实体 $E=\{e_1,e_2,\cdots,e_i,\cdots,e_l\}$,同时判断每个实体 e_i 的标注类别 $T_m=\{t_j,t_{j+1},\cdots,t_k\}$(可以有多个类别)。标注条件使用 $F=E\cdot T\mapsto\{0,1\}$ 表示,$f(e,t)=1$ 表示实体指称 e 标注为类别 t,将实体 e 更新到知识库中。

如财税案例中的一个句子"[2018 年 3 月]$_{时间}$[新丰机械公司]$_{企业}$将公司持有的一幢[写字楼]$_{商业建筑}$[出租]$_{事件}$,[租金]$_{收入}$每年为[3 000 万元]$_{金额}$"。通过实体识别,抽取出句中的实体"写字楼",计算学习函数 f(写字楼,商业建筑)值为 1,将实体"写字楼"标注为"物品/财产/建筑/商业建筑",这些标注信息可以用于税务分析(商业建筑与非商业建筑有不同的税率)。

本文实体识别与标注方法的总体系统框架如

图 1所示。根据领域知识库构建实体标注类型集,并使用远程监督的方法构建用于实体识别与标注的语料库。使用深度学习的方法学习训练语料中领域实体特征,确定新实体边界。通过集成学习的方法对实体类别进行预测,作为相应概念的实例并进行标注,最后根据类别标签将新的实体加入到知识库中。

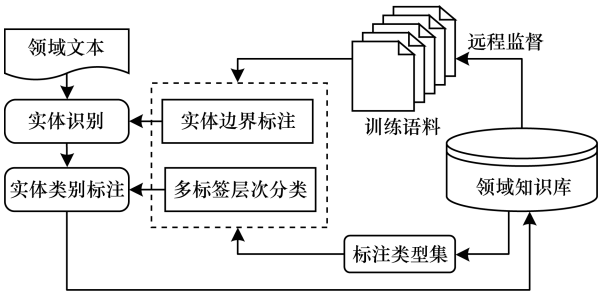

图 1 实体识别与标注方法总体系统框架

Fig. 1 Overall system framework of entity recognition and tagging method

2.2 财税领域知识库

文献[15]使用半自动化方法构建财税领域知识库,构建过程如图 2 所示。由领域专家确定领域中的基本概念和概念间的关系,同时参考现有本体中与财税相关的知识结构,建立顶层本体;使用规则和统计相结合的方法进行术语抽取和关系抽取;通过专家验证将抽取的概念和关系加入到顶层本体中。目前,财税知识库包含1 326个概念、2 326个关系及23 543 个实例。

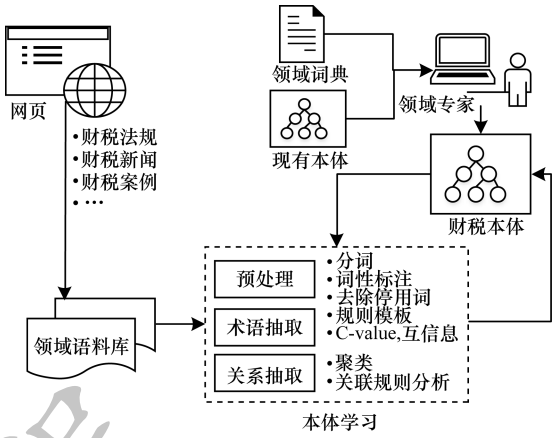

图 2 财税领域知识库构建过程

Fig. 2 Procedure of knowledge base construction in the fiscal and taxation domain

2.3 标注集和语料库构建

2.3.1 实体标注集

由于知识库中的概念多且复杂,有些概念下的实例过于稀少,因此对所有的实体进行类别标注难度较大。为了构建合理的实体标注类型集,本文综合考虑类型的多样性、流行度及其在领域中的重要程度,以保证每个类别都有一定数量的实例。本文选取实例数大于 15 的本体类交给领域专家进行评估,最后确定了 263 个重要类别,并根据知识库的结构进行组织,形成 7 个顶级类别、256 个子类、最大深度为 4 的层次类别集合,标注集示例如图 3 所示。

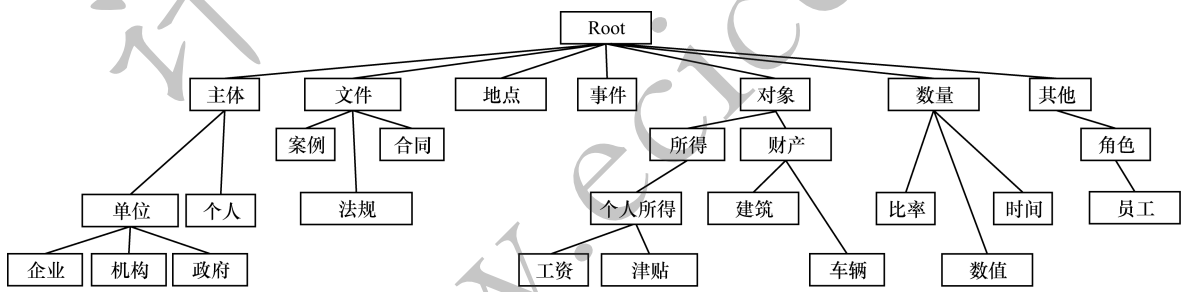

图 3 财税实体标注集示例

Fig. 3 Example of entity tagging set in the fiscal and taxation domain

类别集合中包括 7 个大类:

1)主体(Agent):对客体有认识和实践能力的对象,主要包括单位和个人。

2)物体(Object):任何能够被感知或客观存在的东西,在财税领域主要指征税的对象,如所得、货物、服务、财产、资源等。

3)文件(Document):意识或思想的书面表达形式,财税领域中主要涉及法规、案件、合同、票据、报表等。

4)事件(Event):在特定时间和地点发生的可观察到的事件,如购买、出售、出租、租赁等。

5)地点(Location):描述地理信息,如国家、省、区域等。

6)数量(Quantity):事物的多少、比例、大小、货币等。

7)其他不属于前 6 类的实体,如角色、税种、行业等。

2.3.2 训练集

标注的语料主要为财税案例集,财税案例中一般为某类涉税行为及其处理方法的文本描述,其中包含了大量的领域实体。此外,解释型的法规中也包含了一些重要实体,为了更全面地获取实体信息,本文同时选取了部分解释型法规。

本文标注语料的过程使用文献[16]提出的远程监督方法,该方法会产生一些噪音数据,为了提高训练集的质量,本文使用多种启发式规则对标注语料进行清理。首先针对标注有多个同级类型的实体,删除同级类型标注,只保留其父类型,其次删除标注类型与预定义类型不一致的实体标注,最后删除出

现次数少于设定阈值的实体。

2.4　基于字词特征结合的实体识别

相比于传统的命名实体,中文财税领域的实体识别面临以下挑战:

1)包含的实体数量和种类多。

2)待识别的实体可能会由许多单词修饰,导致实体的边界难以划分。

3)财税领域语言没有统一的命名方式,因此,待识别的实体可能会有多种表述方式,这些实体的特征通过手工方式通常难以准确提取。

基于以上原因,本文使用基于深度学习的策略提取实体特征,结合 CRF 模型对财税领域的实体进行识别。在基于深度学习的实体边界识别研究中,词向量被证明能够在一定程度上提高浅层机器学习方法的效果[12,17-18]。但是,词向量无法对字符级的特征进行很好地表示,原因是中文的字和词都具有一定的语义信息,相同的字组成不同顺序的词语,其语义可能差别很大。因此,研究基于字的信息可以获取更多的实体特征,进而提高实体识别的准确率。但是,基于字的特征由于窗口大小的限制,导致其获取信息的能力有限。另外,中文词语具有特殊的含义,仅使用字的特征也无法高效关联出字词之间的联系。

本文综合考虑实体的字符特征及词特征,对实体边界进行识别,如图 4 所示,模型整体框架共由三部分组成:

1)获取输入句子的词向量表示,对每个词获取其中每个字的向量,字向量再组成词的字向量矩阵,通过卷积神经网络(Convolutional Neural Network,CNN)对字向量矩阵进行卷积和池化,获取每个词的字特征。

2)对每个词的字向量和词向量进行拼接,将拼接结果输入双向长短时记忆(Bidirectional Long Short-Term Memory,BLSTM)网络进行实体识别。

3)由 CRF 层对 BLSTM 层的输出进行解码,得到最优的标记序列。

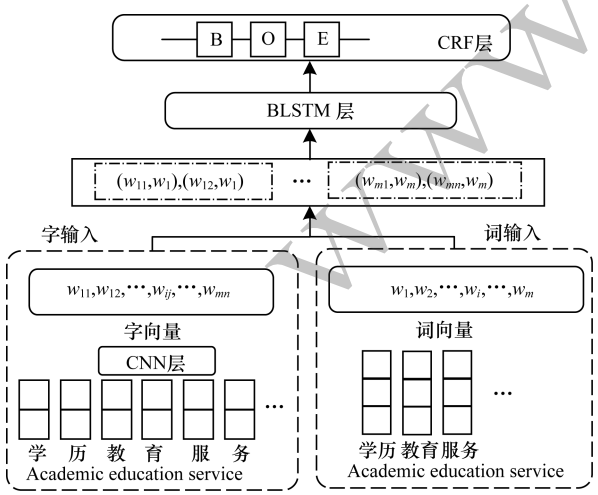

图 4　基于字词特征结合的深度神经网络模型

Fig. 4　Deep neural network model based on the combination of character and word characteristics

CNN 中的卷积层对数据的局部特征具有较好的描述能力,通过池化层可以抽取出局部特征中最具代表性的内容[17]。CNN 的结构主要包括字向量表、卷积层和池化层,如图 5 所示。首先字向量表将词中的字转化成对应的字向量并组成词的字向量矩阵,以长度最大的词为准,在词的左右两端补充占位符(padding)使所有字向量矩阵的大小一致,模型训练时字向量表通过反向传播算法进行更新;然后在卷积层对词的字向量矩阵进行卷积操作提取词的局部特征,卷积核大小为 T(可以提取词周围 T 个词的特征);最后通过池化操作对特征进行压缩获得词的字向量。

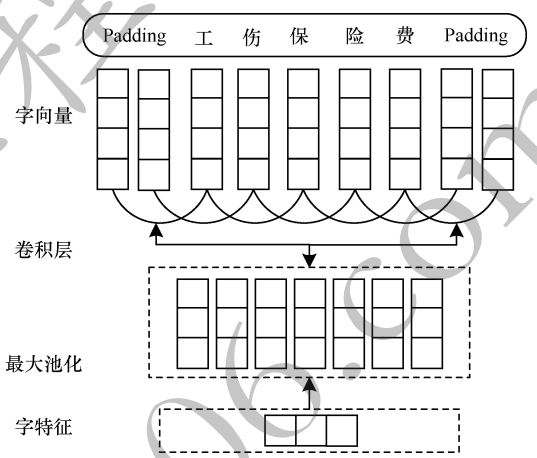

图 5　字级 CNN 模型

Fig. 5　Character-level CNN model

LSTM 模型的门结构可以有选择地保存上下文信息[19],本文利用 BLSTM 网络结构,分别采用顺序和逆序对每个句子进行计算,得到 2 个不同的隐层表示,再将相应时刻输出的结果拼接得到最终的隐层表示。对句子各个位置进行标注时没有利用已经标注过的信息,因此,接入 CRF 层进行标注。CRF 层通过分析相邻标签的关系得到全局最优的标记序列,从而进行句子级的序列标注[20]。对于句子 $S=w_1,w_2,\cdots,w_i,\cdots,w_n$,通过训练得到 BLSTM 层输出大小为 $N\times K$ 的矩阵 $\boldsymbol{P}$,其中,N 为词数,K 为标签种类。设 p_{ij} 为句中第 i 个词第 j 个标签的概率,则模型对句子序列 $y=\{y_1,y_2,\cdots,y_n\}$ 的预测概率为:

$$S(X,y)=\sum_{i=0}^{n}\boldsymbol{A}_{y_i,y_{i+1}}+\sum_{i=1}^{n}\boldsymbol{P}_{i,y_i} \tag{1}$$

其中,$\boldsymbol{A}$ 为转移矩阵,大小为 $K+2$(包含句子开始和结束标记)。整个序列的打分等于各个位置的打分之和,可以利用此前已经标注过的标签为一个位置进行标注。使用 Softmax 归一化句子标记序列为 y 的概率如下:

$$p(y\mid S)=\frac{e^{S(X,y)}}{\sum_{y'\in Y_X}S(X,y')} \tag{2}$$

其中,Y_X 为可能标记的集合。模型训练使用最大化似然函数,则标记序列的似然函数为:

$$\ln p(y|S) = S(X,y) - \ln \sum_{y' \in Y_X} e^{S(X,y')} \quad (3)$$

通过式(3)得到合理的输出序列,然后通过 Viterbi 算法预测输出整体最优路径:

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

在模型训练时,使用 RMSprop 作为优化器,能够取得比随机梯度下降 SGD 模型更快的训练速度(学习率设为 0.002),在 BLSTM 输入和输出部分增加 dropout(值为 0.5),可以减轻过拟合现象[21]。

2.5 基于集成学习的实体标注

对已识别出的实体进行语义类别预测的过程可以看作层次分类问题,分类过程允许有多条标签路径以及部分标签深度(实体的标签路径可以以非叶子类别结束)。

2.5.1 层次分类

层次分类问题是机器学习领域中的一个重要研究课题,该问题的求解方法主要分为平面分类法、局部分类法及全局分类法 3 种[22],各方法的对比如表 1所示。

表 1 层次分类方法对比

Table 1 Comparison of hierarchical classification methods

方法	优势	劣势
平面分类法	分类过程简单,效率高,避免了错误传播问题	没有考虑类别结构,容易产生数据倾斜问题
局部分类法	支持多标签,支持部分标签深度	容易产生阻塞问题、不一致性问题
全局分类法	在训练和测试过程中考虑了层次问题	针对特定分类器增加了分类过程的复杂性

由于本文的实体标注任务需要满足多标签和部分标签深度,为了方便分析各类别的实体特点,本文采用局部分类法获取实体的类别标签。标注过程如图 6 所示,首先为类别层次中除 Root 节点外的每个节点训练一个二分类器(图中用虚线框表示),然后对候选实体进行自上而下的分类,由每个类别上的分类器判断实体是否属于当前类别。

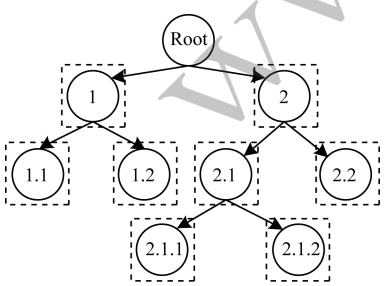

图 6 自上而下的局部分类法

Fig. 6 Top-down local classification method

在层次分类过程中,很多类别实例非常少,即训练样本数量在类别间分布不平衡,因此,容易导致数据过拟合问题。传统的单个分类器方法在解决不平衡分类问题时性能通常会下降,得到的分类器具有偏向性。为了解决该问题,本文使用集成学习方法,根据分类器的差异度选择多个分类器进行集成,以提高分类的准确性。

2.5.2 集成学习

集成学习的目的是根据样本训练多个分类器以对数据集进行预测,解决单个分类器训练数据量小、假设空间小和局部最优解的问题,从而提高整体分类的泛化能力,降低分类误差[23]。

集成学习的方法主要有 Bagging 和 Boosting 2 种,前者的个体分类器间不存在强依赖关系,可以并行执行,后者个体分类器之间有强依赖关系,需要串行执行[6,24]。在实际任务中,由于 Boosting 算法会出现过拟合现象,导致分类结果比单个分类器还差,因此,本文采用 Bagging 算法进行分类器集成。Bagging 是基于数据随机重抽样的分类器构建方法,对于给定训练集 D,其每次训练时根据均匀概率从 D 中有放回地抽取样本 d_i,使用 d_i 训练基分类器 C_i,在采样 T 次后,训练得到 T 个分类器 $C_1, C_2, \cdots, C_t$。对待测样本 x,分别用 T 个分类器进行预测,通过多数投票的方法 $H(x)$ 输出分类结果。由于训练样本为有放回地随机抽取,对训练集中的实例选择没有偏重,增加了集成学习的差异度,从而提高了分类的泛化能力。此外,算法在不稳定点处使用类似于平滑处理的方法,能够提升不稳定分类算法的精度。Bagging 算法详细过程如下:

算法 1 Bagging 算法

输入 training dataset $D = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, Base classifier C, Number of learning rounds T

输出 $H(x) = \underset{y \in Y}{\operatorname{argmax}} \sum_{t=1}^{T} I(h_t(x) = y) \cdot I(y = h_t(x)) = \begin{cases} 1, y = h_t(x) \\ 0, \text{otherwise} \end{cases}$

Iterative process:

For t = 1 to T;

D_t = Bootstrap(D);//a bootstrap sample from D

h(t) = C(D_t)//train a classifier C_t from D_t

传统 Bagging 算法使用一个简单弱分类器,通过处理数据集产生多个差异性分类器。本文对该算法进行改进,将迭代算法中分类器的对象从单个算法扩充到不同的分类算法。分类算法的选择使用了文献[25]中分类器差异性评估的方法,选择支持向量机、逻辑回归及多层感知机 3 种分类器进行集成,对于每个分类器的分类结果采用简单投票法[6]确定最终类别。

2.5.3　特征选择

在分类任务中,特征选取是一个很重要的步骤[26]。对于实体的分类特征,本文参考文献[5,14]中关于实体特征的描述,此外还考虑实体所在句子中的临近动词特征、实体本身的词向量特征以及实体所在文档的主题特征,实体标注特征集如表 2 所示。

表 2　实体标注特征集

Table 2　Feature set of entity tagging

特征	描述	示例
实体	实体本身的字符特征	粮食、白酒
长度	实体的长度大小	4
上下文	窗口大小为 5 的上下文内容	本月、销售、取得、销售额
主题	实体所在文档的主题信息	增值税、消费税
临近动词	句子中距离实体最近的动词	销售、取得
词向量	训练得到的实体的词向量信息	白酒、生产、价格、葡萄酒

3　实验结果与分析

3.1　评估指标

对于实体识别测试,本文使用准确率、召回率及 F 值进行评估,三者的计算方式如下:

$$\text{Precision}=\frac{\text{正确标注的实体数}}{\text{测试集中的实体数}}\times 100\% \tag{5}$$

$$\text{Recall}=\frac{\text{正确标注的实体数}}{\text{标注出的实体数}}\times 100\% \tag{6}$$

$$F=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}\times 100\% \tag{7}$$

对于实体标注结果的评估,本文借鉴层次分类问题中的经典评估指标,即使用 3 种不同粒度(Strict、Macro 及 Micro)的准确率、召回率和 F 值[14]。在实体识别过程中,被错误识别的实体记为错误标注,设 T 为应该被识别出的实体集合,P 为使用本文方法识别出的实体集合。对实体 e,正确的类别标签集为 t_e,预测的标签为 $\hat{t}_e$。上述 3 种粒度的准确率与召回率计算方法分别如下:

1)Strict:当且仅当 $t_e=\hat{t}_e$ 时记为正确标注。

$$P_{\text{strict}}=\frac{\sum_{e\in P\cap T}\delta(t_e=\hat{t}_e)}{|P|},R_{\text{strict}}=\frac{\sum_{e\in P\cap T}\delta(t_e=\hat{t}_e)}{|T|} \tag{8}$$

2)Macro:计算每个实体的准确率及召回率。

$$P_{\text{Macro}}=\frac{1}{|P|}\sum_{e\in P}\frac{|t_e\cap\hat{t}_e|}{|\hat{t}_e|}$$

$$R_{\text{Macro}}=\frac{1}{|T|}\sum_{e\in T}\frac{|t_e\cap\hat{t}_e|}{|\hat{t}_e|} \tag{9}$$

3)Micro:计算总体的准确率与召回率。

$$P_{\text{Micro}}=\frac{\sum_{e\in P}|t_e\cap\hat{t}_e|}{\sum_{e\in P}|t_e|},R_{\text{Micro}}=\frac{\sum_{e\in T}|t_e\cap\hat{t}_e|}{\sum_{e\in T}|t_e|} \tag{10}$$

其中,P 为准确率,R 为召回率。

3.2　实验设置

本文实验使用的实体识别和标注语料库包括了 1 000 个财税法规和 3 000 个财税案例,通过远程监督的方法标注文本中的实体及其类别信息,选取其中 800 个法规及 2 200 个案例作为训练集,然后由领域专家通过标注平台人工标注 200 个法规及 800 个案例作为测试语料。语料基本统计信息如表 3 所示,标注示例如图 7 所示。

表 3　语料统计信息

Table 3　Statistical information of corpus

内容	数量
文档数	1 000
句子数	42 143
实体数	62 830
类别标签数	68 326
一级类别标签数	40 834
二级类别标签数	16 227
三级类别标签数	7 218
四级类别标签数	4 047

语料

[2010 年 5 月]份,[蓝天建筑安装公司]的经营业务如下:销售[建筑装饰材料]销售额[82 万元],取得的[增值税专用发票]的进项税额[10.8 万元]

标注信息

2010 年 5 月:数量/时间

蓝天建筑安装公司:主体/单位/企业

建筑装饰材料:物体/货物

增值税专用发票:文档/票据

82 万元,10.8 万元:数量/金额

图 7　实体标注示例

Fig. 7　Example of entity tagging

3.3　结果分析

3.3.1　实体识别评估

本文在语料处理时使用 BIOES(Begin,Inside,Other,End,Single)方法代替 BIO 来标记实体,这样能够清楚地表示和划分语料中待识别的领域实体边界[27]。深度神经网络模型在整个语料集上的实体识别结果如表 4 所示(W 表示词特征,C 表示字符特征)。

表 4 不同方法的实体识别结果对比

Table 4 Comparison of entity recognition results of different methods %

方法	Precision	Recall	*F* 值
BLSTM(W)	68.34	65.76	67.03
BLSTM + CRF(W)	70.13	69.39	69.75
CNN + BLSTM + CRF(W)	71.67	70.89	71.28
CNN + BLSTM + CRF(C + W)	74.28	72.15	73.20

为了验证 CRF 模块的有效性,将 BLSTM + CRF 与 BLSTM 模型进行对比,从表 4 可以看出,增加 CRF 模块后实体识别的准确率、召回率、*F* 值均有所提高,主要原因是 CRF 能够利用相邻标签的关系对序列进行全局标注,提高对较长及带有修饰词汇的财税实体的识别效果。对比 CNN + BLSTM + CRF 和 BLSTM + CRF 模型发现,CNN 模块对实体抽取结果也有一定程度的提高,这是因为 CNN 模块抽取的字向量能够表示实体的形态特征。最后对比基于词向量的方法和基于字词向量相结合的方法,结果表明,相比基于词向量的方法,字向量与词向量相结合的方法在准确率、召回率和 *F* 值上分别提高了 2.61%、1.26% 和 1.92%,这是由于字词向量相结合的方法将实体的词特征和字符特征进行组合,得到了更丰富的特征表示,对于长度较长、带有修饰词汇的实体以及罕见实体(如“长期股权投资”),其识别性能较高。

利用基于词向量的方法及基于字词向量相结合的方法,在财税领域对各类别中的实体(包括主体、文件、物体、事件、地点、数量及其他类别)进行识别,识别结果如表 5 所示。

表 5 各类别中的实体识别结果

Table 5 Entity recognition results of each category

类型	方法	数量	Precision/%	Recall/%	*F* 值/%
主体	W	1 824	72.14	68.21	70.12
	C + W	2 115	75.78	71.72	73.69
文件	W	1 257	82.23	83.26	82.74
	C + W	1 312	85.06	85.38	85.22
物体	W	4 328	64.21	63.13	63.66
	C + W	5 063	68.33	65.27	66.76
事件	W	486	69.11	66.60	67.83
	C + W	571	69.48	67.35	68.40
地点	W	1 326	85.62	83.91	84.76
	C + W	1 411	87.46	84.47	85.94
数量	W	2 121	86.35	84.71	85.52
	C + W	2 218	87.83	85.82	86.81
其他	W	621	71.38	65.40	68.26
	C + W	684	72.17	69.64	70.88

从表 5 可以看出,主体、文件、地点和数量类实体的识别准确率、召回率和 *F* 值相对较高,而物体类实体识别效果较差,这是因为物体类实体比其他类别中的实体更加复杂,而且有些子类中的实例相对较少。在物体类实体中,相比词向量方法,字词向量相结合的方法性能提高最明显,原因是物体类别中的实体平均长度较长,字向量的加入提供了更多子词的特征。此外,本文进一步分析实体识别错误的情况,发现错误的原因很大程度上是由于训练语料中缺乏对相关类型实体的标注或者标注错误。因此,下一步考虑使用主动学习的方法,发现错误率较高的实体类型中的实例即交由专家标注,以提高实体识别的整体效率。

3.3.2 实体标注评估

在整个语料库中,分别利用本文集成学习方法、逻辑回归、支持向量机及多层感知机方法对实体进行标注。其中,逻辑回归的正则参数设为 0.3,支持向量机的参数 *C* 设为 0.05,多层感知机的隐含层大小设为 100,集成学习使用的迭代次数设为 30。训练过程中类别标签下的实例为正样本,其兄弟节点的实例为负样本,测试结果(*F* 值)如表 6 所示。

表 6 不同方法的实体标注测试结果

Table 6 Testing results of entity tagging of different methods %

方法	Strict	Macro	Micro
逻辑回归	45.41	53.76	56.13
支持向量机	49.33	57.69	61.76
多层感知机	44.67	53.19	55.72
本文集成学习	53.52	64.80	66.17

从表 6 可以看出,相比单个分类器的方法,集成学习在 Strict、Marco 及 Micro 上的 *F* 值均有明显提高。在各类方法中,Strict 上的 *F* 值相对其他两项指标明显偏低,这是因为在上一步实体识别的过程中产生了很多错误的候选实体。Macro 上的 *F* 值比 Micro 略低,原因是实体类型层结构中部分底层的类别实例较少,产生了长尾效应。进一步对标注错误的情况进行分析发现,标注错误的情况比较集中,某些类别的标注准确率较低,很大程度上是由于这些类别中训练数据不足。

各不同层级类别标签的标注结果如图 8 所示,为了简单起见,本文仅统计标注结果 Micro 上的 Precision、Recall、*F* 值,使用的方法为本文集成学习方法。从图 8 可以看出,层次越深的类别,实体标注的效果越差,原因是高级别的类别具有更多的实例,类别之间的区分更明确,而层次结构较低的类别实例较少,实体的特征区别不明显,甚至对于有些类别,人工标注可能也难以判断。

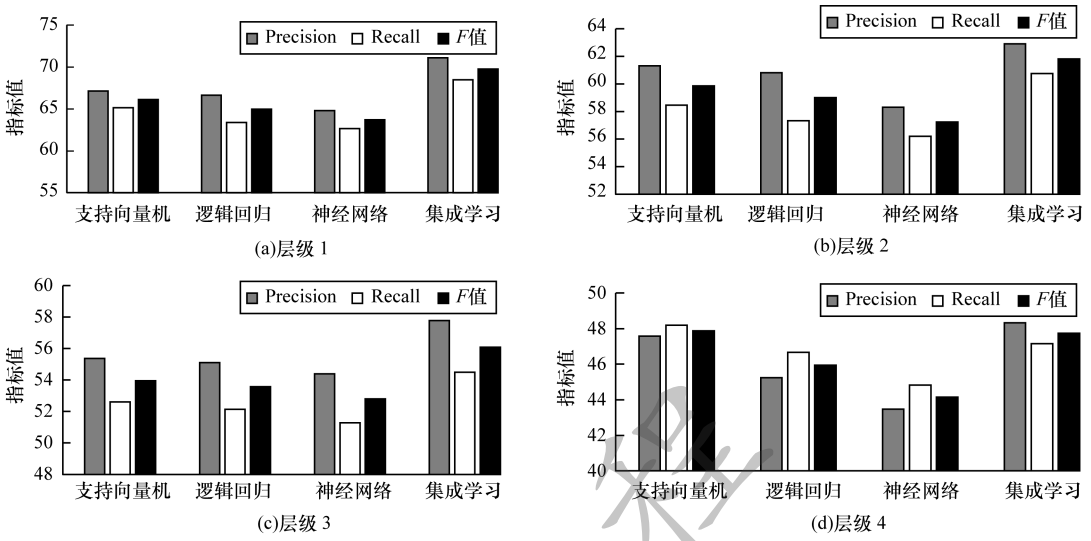

图 8　各层级实体标注实验结果

Fig. 8　Experimental results of entity tagging of each hierarchy

本文对顶级类别中实体标注的实验结果进行统计和分析,结果如图 9 所示。从图 9 可以看出,在数量和地点类别中,标注的准确率、召回率及 F 值明显高于其他类别,这是因为这些类别中的实体特征更为明显。相对地,物体类别的标注效果最差,原因是该类别中包含的子类更多,种类更为复杂。因此,下一步需要根据物体类别中的实体来选择更有效的特征进行分析测试。

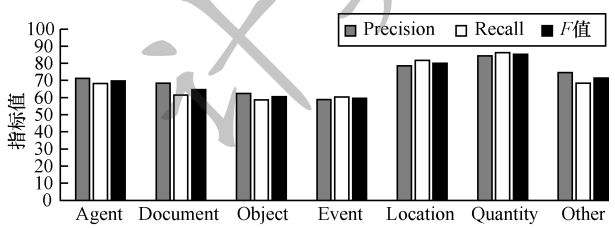

图 9　各顶级类别中的实体标注实验结果

Fig. 9　Experimental results of entity tagging of each top level

本文通过从特征集中删除某个特征来判断该特征对实体类别标注结果的影响,评估指标同样使用 Micro 上的 Precision、Recall、F 值,标注方法为本文集成学习方法。从表 7 可以看出,去除主题特征、临近动词特征或词向量特征后标注结果的准确率、召回率及 F 值均有不同程度的下降,即主题特征对标注结果具有影响,特别是对于物体、主体类别,原因是一般特征主题都有相应的纳税人和征税对象类型。临近动词的影响主要体现在和特定类型的实体之间一般有固定的关联关系(例如,在动词“签订(签字)”后的实体更可能被分类为“合同”类型)。通过实验结果还可以看出,词向量特征在预测不常见实体类型时的作用比较明显。

表 7　各类实体特征测试结果

Table 7　Testing results of each type of entity feature　%

特征	Precision	Recall	F 值
所有特征	68.26	65.84	67.03
主题特征	63.36	62.53	62.94
临近动词特征	64.75	63.41	64.07
词向量特征	62.18	64.27	63.20

4　结束语

本文对财税领域的实体识别和标注方法进行研究,根据财税知识库定义一组领域实体标注集,使用远程监督方法构建领域语料库。通过深度学习方法对中文财税领域的实体识别进行测试,采用基于字向量与词向量相结合的深度神经网络模型解决词向量模型无法描述局部特征的问题。在此基础上,使用层次分类的方法预测候选实体类别,并提出一种集成学习方法解决数据不平衡和单个分类器偏见的问题。实验结果表明,该方法具有较高的准确率。下一步将研究和优化集成学习方法,在实体标注的集成方法中测试更多基分类器,如决策树、朴素贝叶斯、k-最近邻算法等,进一步优化分类效果,同时研究使用联合学习方法将 2 个子任务合并成一个序列标注的问题。

参考文献

[1] FUJITA H, ALI M, SELAMAT A, et al. Trends in applied knowledge-based systems and data science[M]. Berlin, Germany: Springer, 2016.

[2] HOU Mengwei, WEI Rong, LU Liang, et al. Research review of knowledge graph and its application in medical domain [J]. Journal of Computer Research and Development, 2018, 55(12): 2587-2599. (in Chinese)
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.

[3] NIU F, ZHANG C, RE C, et al. DeepDive: Web-scale knowledge-base construction using statistical learning and inference[J]. VLDS, 2012, 1(12): 25-28.

[4] MITCHELL T, COHEN W, HRUSCHKA E, et al. Never-ending learning[J]. Communications of the ACM, 2018, 61(5): 103-115.

[5] ABHISHEK A. FgER: fine-grained entity recognition[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. California, USA: AAAI Press, 2018: 8008-8009.

[6] ZHOU Zhihua. Ensemble methods: foundations and algorithms[M]. [S.l.]: CRC Press, 2012.

[7] JIA Yidong, XU Weiran, QIN Pengda, et al. Fine-grained entity typing for knowledge base completion[C]//Proceedings of 2016 IEEE International Conference on Network Infrastructure and Digital Content. Washington D. C., USA: IEEE Press, 2016: 361-365.

[8] GILLICK D, LAZIC N, GANCHEV K, et al. Context-dependent fine-grained entity type tagging[EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1412.1820.pdf.

[9] LEE C, HWANG Y G, OH H J, et al. Fine-grained named entity recognition using conditional random fields for question answering[C]//Proceedings of Lecture Notes in Computer Science. Berlin, Germany: Springer, 2006: 581-587.

[10] LING X, WELD D. Fine-grained entity recognition[C]// Proceeding of the Association for the Advancement of Artificial Intelligence. California, USA: AAAI Press, 2012: 1-7.

[11] LIU Liu, WANG Dongbo. A review on named entity recognition[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329-340. (in Chinese)
刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[12] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF[EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1603.01354.pdf.

[13] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends indeep learning based natural language processing[J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.

[14] YOSEF M A, BAUER S, HOFFART J, et al. Hyena: hierarchical type classification for entity names[C]// Proceedings of the 24th International Conference on Computational Linguistics. Washington D. C., USA: IEEE Press, 2012: 1361-1370.

[15] QIU Y, CHENG L, ALGHAZZAWI D. Towards a semi-automatic method for building Chinese tax domain ontology[C]//Proceedings of the 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Washington D. C., USA: IEEE Press, 2017: 2530-2539.

[16] REN Xiang, HE Wenqi, QU Meng, et al. Afet: automatic fine-grained entity typing by hierarchical partial-label embedding[C]//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2016: 1369-1378.

[17] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Computer Science, 2015, 11(1): 1-14.

[18] Maimaitlayifu, SILAMU Wushouer, MUHETAER Palidan, et al. Uyghur named entity recognition based on BiLSTM-CNN-CRF model[J]. Computer Engineering, 2018, 44(8): 230-236. (in Chinese)
买买提阿依甫, 吾守尔·斯拉木, 帕丽旦·木合塔尔, 等. 基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别[J]. 计算机工程, 2018, 44(8): 230-236.

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[20] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, 2015, 8: 1-10.

[21] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 56(15): 1929-1958.

[22] SILLA C, FREITAS A. A survey of hierarchical classification across different application domains[J]. Data Mining & Knowledge Discovery, 2011, 22(1): 31-72.

[23] ZHOU Gang, GUO Fuliang. Research on ensemble learning[J]. Computing Technology and Automation, 2018, 37(4): 148-153. (in Chinese)
周钢, 郭福亮. 集成学习方法研究[J]. 计算技术与自动化, 2018, 37(4): 148-153.

[24] ROKACH L. Ensemble-based classifiers[J]. Artificial Intelligence Review, 2010, 33(1/2): 1-39.

[25] YANG Chun, YIN Xucheng, HAO Hongwei, et al. Classifier ensemble with diversity: effectiveness analysis and ensemble optimization[J]. Acta Automatica Sinica, 2014, 40(4): 660-674. (in Chinese)
杨春, 殷绪成, 郝红卫, 等. 基于差异性的分类器集成: 有效性分析及优化集成[J]. 自动化学报, 2014, 40(4): 660-674.

[26] TRIGUERO I, VENS C. Labelling strategies for hierarchical multi-label classification techniques[J]. Pattern Recognition, 2016, 56(8): 170-183.

[27] GOYAL A, GUPTA V, KUMAR M. Recent named entity recognition and classification techniques: a systematic review[J]. Computer Science Review, 2018, 29(8): 21-43.

编辑 吴云芳