税收优惠政策关键要素抽取与可视化分析

关海山[1,2]，郑玉龙[1,2]，魏笔凡[2,3]，张泽民[1,2]，岳浩[1,2]，师斌[2,4]，董博[2,4]
1. 西安交通大学软件学院，陕西 西安 710049；
2. 陕西省天地网技术重点实验室，陕西 西安 710049；
3. 西安交通大学继续教育学院，陕西 西安 710049；
4. 西安交通大学计算机科学与技术学院，陕西 西安 710049

摘要
随着税收优惠政策数量的迅速增加，纳税人面对海量的税收优惠政策难以快速定位与自身相关的优惠内容，导致许多纳税人没有享受到应该享受的优惠政策。基于预训练语言模型BERT与规则处理相结合的方法实现了对税收优惠政策法规的表示、关键要素抽取和税收优惠的可视化查询，使纳税人可以快速准确地定位与自身相关的税收优惠信息，并对结果进行可视化展示。实验结果表明，关键要素抽取性能优越，税收优惠政策查询快速直观，可有效缓解海量税收优惠信息过载。

关键词
税收优惠政策；预训练语言模型；信息抽取；可视化

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022035

Extraction and visualization analysis of key elements of tax preferential policies

GUAN Haishan[1,2], ZHENG Yulong[1,2], WEI Bifan[2,3], ZHANG Zemin[1,2], YUE Hao[1,2], SHI Bin[2,4], DONG Bo[2,4]
1. School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
2. Shaanxi Key Laboratory of Satellite & Terrestrial Network Technology R&D, Xi'an 710049, China
3. School of Continuing Education, Xi'an Jiaotong University, Xi'an 710049, China
4. School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

Abstract
With the rapid increase in the number of preferential tax policies, taxpayers face a large number of preferential tax policies, and it is difficult for taxpayers to quickly locate the preferential content related to them. As a result, many taxpayers do not enjoy the preferential policies they should enjoy. Based on the combination of pre-training language model BERT and rule processing, the visualization was realized of the characterization of preferential tax policies and regulations, the extraction of key elements, and the visual query of tax incentives, so that taxpayers can intuitively

and quickly locate tax incentives related to themselves, and visualize the results. The experimental results show that the extraction performance of key elements is superior, and the query of preferential tax policies is quick and intuitive, which can effectively alleviate the problem of massive tax preferential information overload.

Key words
preferential tax policy, pre-trained language model, information extraction, visualization

0 引言

税收优惠政策是指税法对某些纳税人和征税对象给予鼓励和照顾的一种特殊规定，是国家利用税收调节经济的具体手段。国家通过税收优惠政策可以扶持某些特殊地区、产业、企业和产品的发展，促进产业结构的调整和社会经济的协调发展，保证了企业的竞争力和存活力，并且对就业和再就业具有极大的积极影响。税收优惠政策的形式各种各样，包括税额减免、税基扣除、税率降低等。税收优惠政策的范围越广、差别越大、方式越多、内容越丰富，纳税人税收筹划的空间就越大、节减税收的合理方式就越多，因此纳税人可合法利用税收优惠政策来减轻自身的税收压力。

为了“减税降负”“精准施策”以及推进“放管服”等改革措施，税务主管部门近年来推出了大量不同类型的税收优惠政策。这些税收优惠政策主要通过专题讲座、纳税教育辅导以及网站政策公告等方式进行宣传和推广，时效性差、覆盖面小。纳税人需要花费大量的时间跟踪税收优惠政策的发布，快速从海量税收优惠政策中查找并定位与自身相关的优惠信息变得越来越困难，导致许多纳税人没有享受到应该享受的优惠，甚至有些纳税人不清楚哪些优惠政策与自己相关。随着互联网的快速发展，网络数据呈现出大规模、多元化、组织结构松散等特点。税收优惠信息也难以避免这种情况，多源、异构导致的信息碎片化等问题[1]给纳税人获取有效的税收优惠信息造成了困扰。

国家税务总局在《关于进一步深化税收征管改革的意见》中强调，要优化以满足纳税人正当需求和维护合法权益为中心的纳税服务，构建更加方便、快捷、高效的纳税服务机制[2-3]。目前少数研究者希望通过大数据技术等前沿技术实行税收优惠政策的“直达快享”，但是税务大数据推荐技术需要从多个维度分析纳税人的过程信息和行为信息，而目前的税务信息系统还无法提供这些信息。此外，为了保证大数据推荐技术的质量，需要集成税务领域中大量的业务系统，但是各个系统提供的原始数据往往存在标准不统一、一致性低、规范性差等问题，需要进行海量数据的清洗、转换、对碰等预处理工作，工作量大、错误率高。利用少量数据预测大量未知信息则可能产生过拟合等风险[1,4]。

为此，本文基于深度学习与特征规则联合抽取方法构建了一个税收优惠法规可视化分析系统，该系统的贡献主要体现在以下两点：

- 根据税务专家的经验知识，制作税收优惠关键要素数据模板，提供了一种针对税收优惠政策的结构化数据抽取方法，解决了税收优惠政策信息碎片化等问题，完成了复杂税收优惠信息到结构化信息的转换；
- 基于税收优惠关键要素数据模板，

设计了以纳税人为主体的径向图可视化查询方式，解决了纳税人在面对海量的税收优惠政策时，无法快速定位与自身相关的优惠内容的问题。

该系统的构建主要有3个步骤：税收优惠主题构建、税收优惠分面识别、税收优惠查询与可视化。

（1）税收优惠主题构建

- 文本分割：对税收优惠政策进行文本分割处理。根据大量观察与统计，税收优惠政策的文件表现形式一般为由若干个条款组成的完整文档，每个条款都描述了一些与其他条款不同的税收信息。因此根据优惠文档的特征设计出文本分割算法，把一个税收优惠政策文档分割为若干个税收优惠条款，得到一个由若干个条款组成的集合。
- 税收条款的优惠信息识别：将单个税收优惠政策处理为若干个条款后，并非所有条款的内容都包含与税收优惠相关的信息，因此构建一个深度学习的分类模型，识别出与税收优惠相关的条款。

（2）税收优惠分面识别

制作税收优惠政策关键要素数据模板，该模板包含条款内容、享受主体、标题、文号、减免方式、减免类型、税种、政策类型和有效期限9个关键要素。根据不同的关键要素构建不同类型的模型任务对其内容进行识别和抽取，然后使用关系型数据库将抽取的知识进行存储，为查询与可视化提供数据支撑。

（3）税收优惠查询与可视化

面对海量税收优惠政策文件，纳税人难以精准检索到相关税收优惠内容，且难以直接了解税收优惠的重要信息。因此，该系统设计了税收优惠政策查询与可视化的功能模块。当用户输入税收优惠政策的享受主体后，就能快速查询到该享受主体以及与其相似的享受主体相关的税收优惠政策，并以径向图的方式展示，显示每个政策条款的关键要素内容，提高政策条款的易读性。

1 相关工作

近年来税务领域的相关工作侧重于偷税漏税检测、发票虚开检测、金融欺诈识别等，文本信息抽取方面的工作较少。因此，本节将从两方面进行介绍，一是针对某一特定领域的文本信息抽取工作，二是可视化布局的相关工作。

1.1 特定领域内的文本信息抽取工作

在特定领域内进行文本信息抽取的工作已有许多。针对特定领域中的语料个性化、训练数据稀缺等问题，如何进行文本信息抽取工作是研究者一直关心的问题。Zhang R X等人[5]对少量监管文件和物业租赁协议文档进行人工注释，利用这些文档对BERT（bidirectional encoder representations from transformers）模型进行微调，之后成功利用该模型从这两种不同类型的商业文档中提取结构化实体，并将成果展示在一个端到端云平台，允许用户上传文档并检查模型的结果，说明少量特定领域的注释数据足以微调BERT模型，实现具有一定准确度的元素内容的提取。Nguyen M T等人[6]在BERT模型上叠加卷积神经网络（convolutional neural network，CNN）层完成了迁移学习，基于Transformers开发了原型产品AURORA，该系统解决了在训练样本数量有限的情况下，从特定领域中提取结构化信息的问题。Friedrich A等人[7]针对材料科学领域提出了3个信息提取任务：实验描述句子的检测、实体识别和输入以及与实验相关的数值的识别，针对这些任务，

他们标注了一个新的语料库，使用不同的模型进行信息抽取的对比工作，实验发现BERT模型的性能优于其他模型的性能，同时他们使用BERT+BiLSTM（双向长短期记忆网络）的组合以应对更加复杂的挑战。Zeghdaoui M W等人[8]提出了一种基于CNN结合长短期记忆（long short-term memory，LSTM）神经网络的医学文本分类模型，CNN-LSTM模型使用通过FastText计算的词向量来实现最高准确度，获得了较好的结果。

1.2 可视化布局

如何合理地将与纳税人相关的优惠信息可视化，并通过简单直观的方式进行展示，是一个值得思考的问题。Brandes U等人[9-10]提出，中心性是图分析中一个重要的研究内容，它量化了节点在图结构中的重要性，因此径向布局是一种直观地表达节点间相对重要性的有效方法。之后他们又提出了一种新型的径向布局，该方法是基于应力最小化的扩展，其加权方案在优化过程中逐渐对中间布局施加径向约束。Raj M等人[11]提出了一种新的无向图布局方法，将顶点约束在一组闭合的曲线上，这种布局可以很好地显示图的中心性和顶点距离信息，同时提供了一种可视化策略证明了布局方法的有效性。Fenu G等人[12]在社交网络、YouTube、Wikipedia上使用了径向布局来表示用户与特定对象的匹配关系，认为简单而有效的可视化状态可以给用户带来不同的好处。Bostock M等人[13-14]提出了ProtoVis和D3.js框架，ProtoVis可以将数据直接映射到可视元素，使设计者无须计算细节即可实现可视化；D3.js可以将输入数据绑定到任意的文档元素中，通过动态转换修改内容。Li D Q等人[15]提出了Echart可视化框架，它是一个开源的、基于Web的、跨平台的框架，具有简单易用、交互内容丰富以及高性能的特点，它的核心是一套声明式可视化设计语言，设计者可以自定义内置图表类型。

本文的主要工作是抽取税务领域中的一些关键信息，通过实验对比将性能较好的BERT模型作为核心，针对不同的信息抽取任务采用不同的处理方式，实现对税收优惠关键信息的抽取，并采取径向图布局的方法进行可视化展示。

2 系统概述

2.1 系统结构框架

该系统包含两个概念定义：税收优惠主题和税收优惠条款分面。将每个税收优惠政策文档看作一个独立的集合，用N表示，将每个文档内部包含的各个优惠条款看作最小的不可分割的元素，用C表示。定义一个集合N由若干个元素C组成，表示为$N=\{C_1,C_2,\cdots,C_n\}$，如果C_i包含了税收优惠的相关内容，则称C_i为一个税收优惠主题。根据税务专家经验，制作税收优惠政策的数据结构模板。该模板包括条款内容、享受主体、标题、文号、减免方式、减免类型、税种、政策类型和有效期限9个关键要素，这些关键要素可以有效地对税收优惠文档的重要内容进行表示。其中，一个关键要素就是税收优惠条款的一个分面，每个元素C_i都由这9个分面组成。最终的结构为一个税收优惠政策文档包含一个或多个主题，每个主题具有9个分面，每个分面都对应一个关键要素内容。

图1所示为税收优惠法规可视化系统3个模块的框架。每个模块的功能和特性描述如下。

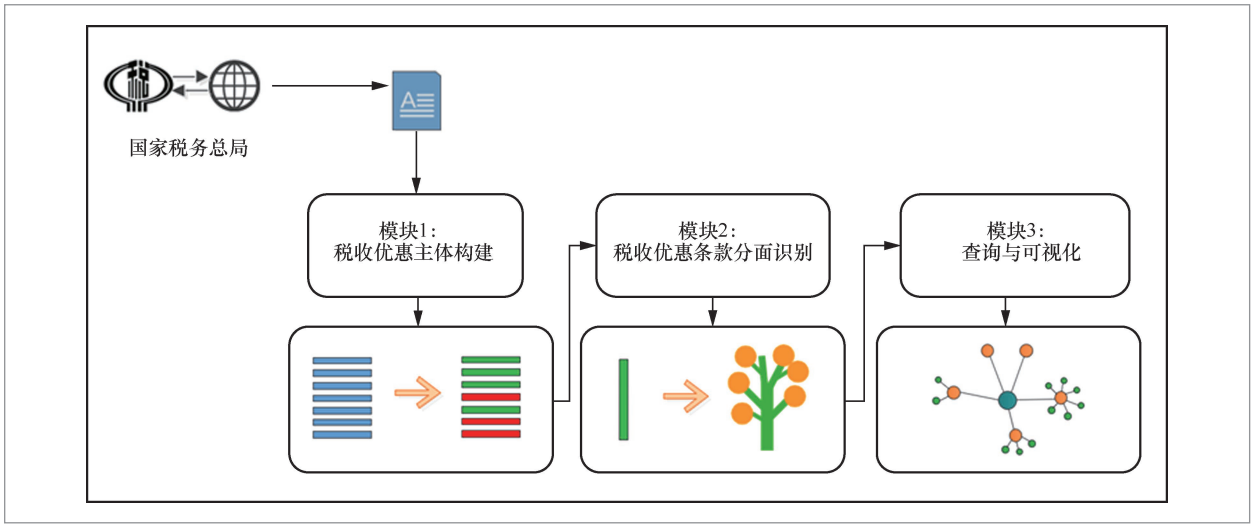

图1 系统框架

为了保证数据源的权威性以及准确性，将国家税务总局官方网站以及各省市地方分局官方网站作为本系统的数据来源。使用Python爬虫技术的Requests库和BeautifulSoup库进行页面文档的全面解析，过滤除文档自身内容以外的不必要元素，以保证数据源的质量。

模块 1：税收优惠主题构建。根据税务专家系统的先验知识，对大量税收优惠政策文档的结构、特征进行归纳和总结。通过特征提取，定位文档关键位置，使用基于规则的方法，设计了针对税收优惠政策文档的文本切割算法，该算法可以将税收优惠政策文档分割为若干条以单个条款为最小文本单位的文本序列集合。最后将每个税收优惠文档形式化为一个独立的集合，用 N 表示。将文档内部的各个优惠条款看作最小且不可分割的元素，用 C_i 表示。定义一个集合 N 由若干个元素 C_i 组成，表示为 $N=\{C_1,C_2,\cdots,C_n\}$。通过深度学习技术构建的算法模型对集合 N 中的每个主题进行识别，得到集合 N 中含有税收优惠政策的主题 C_i，形成新的集合 $T=\{C_i,\cdots,C_m\}$。

模块2：税收优惠条款分面识别。对模块1中集合T的元素进行处理，使用深度学习与规则处理相结合的方法对每个元素进行识别和抽取，使得每个主题都包含9个税收优惠条款分面。该模块的输入为经过模块1处理后得到的集合T，输出为每个条款的各个关键要素信息。

模块3：税收优惠查询和可视化。根据税收优惠关键要素数据模板，设计分类查询以及相应的可视化算法，实现系统的查询与可视化功能。

在该系统中，数据源的预处理部分简单利用了爬虫程序和基于规则的算法解析，因此不进一步描述这些算法的详细实现。

2.2 税收优惠主题构建

首先，对大量税收优惠文档结构、特征进行归纳和总结，通过提取特征、定位文档关键位置的索引，使用基于规则的方法设计文本切割算法，把一个文本分割成若干条款，如图2所示。

之后，对分割后的条款进行数据标注，标记该条款是否包含与税收优惠政策相关的内容，如果包含，则标记为1，否

图 2　税收优惠条款分割示例

则标记为0；然后使用深度学习模型学习带有标记的样本。本系统采用性能较好的BERT模型，BERT模型是一种基于Transformer的Encoder结构的预训练语言模型，通过海量的文本数据训练掩码语言模型（masked language model，MLM）和下一句预测（next sentence prediction，NSP）任务，使BERT模型可以学习更深层的语义信息[16]。在经过预训练的BERT模型上进行微调，可以使一些下游应用表现出更好的效果。

图3所示为税收优惠主题识别模型结构，具体步骤如下。

步骤1：把输入的条款转换为字符级别的序列。如**图3**所示，设置BERT模型可处理的最大序列长度为maxlen，加上首位CLS符号，故可处理的条款最大长度为maxlen-1。对于超出最大长度的输入条款，根据文本的结构特征，优先处理句子的头部和尾部，即将前0.25×maxlen个字符和后0.75×maxlen个字符作为模型输入；对于长度小于maxlen-1的输入文本，填充空字符，后文采取同样的处理方式，不再赘述。

步骤2：序列首增加CLS符号，生成序列表示。

$$\boldsymbol{S}=[\mathrm{cls},t_1,t_2,\cdots,t_{\mathrm{maxlen}-1}] \tag{1}$$

$$\mathrm{input}=V_{\mathrm{lookup}}(\boldsymbol{S}) \tag{2}$$

$$\boldsymbol{E}=\mathrm{Word\_Embedding}(\mathrm{input}) \tag{3}$$

其中，$\boldsymbol{S}$表示输入序列，V表示词表，除了涉及全部字符外，还包括特殊口令CLS、SEP、UNK、PAD和MASK，V_{lookup}是指在词表V中寻找字符的编号，input表示$\boldsymbol{S}$根据词表中的编号计算出的序列。Word_Embedding指将字符映射为词嵌入向量，结果$\boldsymbol{E}$为输入序列的嵌入向量，计算过程是$\mathrm{input}\times W^{\mathrm{e}}$，$W^{\mathrm{e}}$表示计算结果$\boldsymbol{E}$的权重参数，随机初始化其值，在训练过程中根据梯度更新W^{e}。

步骤3：使用BERT对序列嵌入进行特征提取。

①字向量与位置编码：

$$\boldsymbol{P}=\mathrm{Position\_Embedding}(\mathrm{input}) \tag{4}$$

$$\mathrm{Position\_Embedding}=\mathrm{pos}(\mathrm{input})\times W^{\mathrm{p}} \tag{5}$$

根据式（4）计算位置嵌入$\boldsymbol{P}$，式（5）中pos(input)指获得字符在序列中的位置，W^{p}表示计算结果$\boldsymbol{P}$的权重参数。

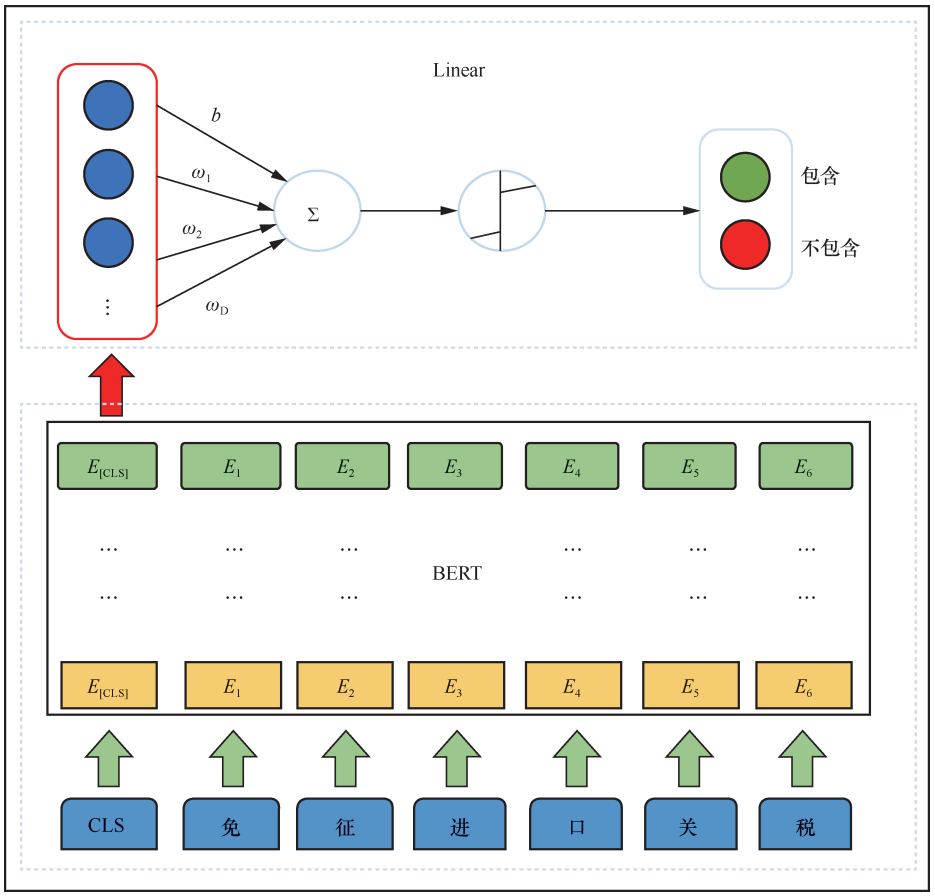

图 3 税收优惠主题识别模型结构

②计算：

$$\boldsymbol{X}=\boldsymbol{E}+\boldsymbol{P} \tag{6}$$

其中，$\boldsymbol{X}$为字符嵌入向量$\boldsymbol{E}$与位置嵌入向量$\boldsymbol{P}$之和。

③自注意力机制：

$$\boldsymbol{Q}=W^{q}\boldsymbol{X} \tag{7}$$

$$\boldsymbol{K}=W^{k}\boldsymbol{X} \tag{8}$$

$$\boldsymbol{V}=W^{v}\boldsymbol{X} \tag{9}$$

$$\boldsymbol{Z}=\mathrm{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\top}/\sqrt{dk}\right)\boldsymbol{V} \tag{10}$$

其中，$\boldsymbol{Q}$为查询矩阵，$\boldsymbol{K}$为键矩阵，$\boldsymbol{V}$为值矩阵，$\boldsymbol{Z}$为自注意力矩阵，W^{q}、W^{k}、W^{v}分别为权重参数，其值进行随机初始化。

④自注意力残差连接与归一化

定义归一化函数：

$$\mathrm{LayNorm}(x)=\frac{x_{ij}-\mu_j}{\sqrt{\sigma_j^2}} \tag{11}$$

计算：

$$\boldsymbol{X}_{\mathrm{attention}}=\boldsymbol{X}+\boldsymbol{Z} \tag{12}$$

$$\boldsymbol{X}_{\mathrm{attention}}=\mathrm{LayNorm}(\boldsymbol{X}_{\mathrm{attention}}) \tag{13}$$

⑤前馈残差连接与归一化：

$$\boldsymbol{X}_{\mathrm{hidden}}=\mathrm{Linear}\left(\mathrm{ReLU}\left(\mathrm{Linear}(\boldsymbol{X}_{\mathrm{attention}})\right)\right) \tag{14}$$

$$\boldsymbol{X}_{\mathrm{hidden}}=\boldsymbol{X}_{\mathrm{hidden}}+\boldsymbol{X}_{\mathrm{attention}} \tag{15}$$

$$\boldsymbol{X}_{\mathrm{hidden}}=\mathrm{LayNorm}(X_{\mathrm{hidden}}) \tag{16}$$

其中，$\boldsymbol{X}_{\mathrm{attention}}$表示自注意力分数，$\boldsymbol{X}_{\mathrm{hidden}}$表示输入序列的隐藏状态。此时，文本的深

层语义特征提取全部完成，为了方便描述，后文统一用$\boldsymbol{X}_{\text{hidden}n}$=BERT（$\boldsymbol{S}$）表示BERT对序列嵌入进行特征提取。

步骤4：使用全链接层将隐藏层第1个位置（CLS对应的特征向量）进行特征提取。此向量包括整句的所有语义信息，全连接层将CLS特征向量维度降至标签个数t。

$$\boldsymbol{Y}=\text{Linear}_t\left(\boldsymbol{X}_{\text{hidden}}[0]\right) \quad (17)$$

步骤5：最后使用Softmax分类器计算相应的标签，$\boldsymbol{Y}$为最终输出结果，即预测标签。

$$\boldsymbol{Y}=\text{Softmax}(\boldsymbol{Y}) \quad (18)$$

由于BERT模型的输入有最大长度限制，为了得到更好的分类结果，对于超出最大长度的条款，按句号切割后分别作为模型的输入，把模型输出的多个结果集成起来作为该条款的分类结果。**图4**展示了税收优惠主题识别的示例，其中第1个和第2个条款包含与税收优惠相关的内容，第3个条款则不包含。

2.3 税收优惠条款分面识别

根据税务专家的经验，在单个条款中人们关心的主要内容和税收优惠关键要素见**表1**，笔者分别以不同的形式对数据进行标注，使用不同的方法和模型进行处理。

2.3.1 享受主体识别

从一个样本序列中识别出享受主体字段，这是一种典型的序列标注任务。例如在“一、自2015年1月1日起至2016年12月31日止，对物流企业自有的（包括自用和出租）大宗商品仓储设施用地，减按所属土地等级适用税额标准的50%计征城镇土地使用税。”这个条款中，“物流企业”是享受主体。把这个样本按照字符顺序拆分成一系列汉字，每个字符都拥有标签，标签类型为“BIO”形式，之后模型需要给出每个字符的标签类型，最终识别为BI标签的字符被认为是享受主体。**图5**所示为享受主体识别模型结构，具体步骤如下。

步骤1：把输入的条款转换为字符级别的序列。

步骤2：序列前端增加CLS符号，生成序列的向量表示。

$$\boldsymbol{S}=[\text{cls},t_1,t_2,\cdots,t_{\text{maxlen}-1}] \quad (19)$$

步骤3：使用BERT对序列嵌入进行特征提取。

图 4 税收优惠主题识别示例

表 1 税收优惠关键要素描述

关键要素	描述	任务处理方式
条款内容	即税收优惠主题，指包含税收优惠信息的条款内容	文本分类
享受主体	描述享受优惠的主体对象	序列标注
税种	优惠涉及的税种	文本多标签分类
减免类型	优惠的减免类型	文本分类
减免方式	优惠的方式	文本分类
政策类型	该优惠条款属于何种政策类型	文本分类
有效期限	优惠条款的时间期限	基于规则
标题	优惠条款所属公告标题	基于规则
文号	优惠条款所属公告文号	基于规则

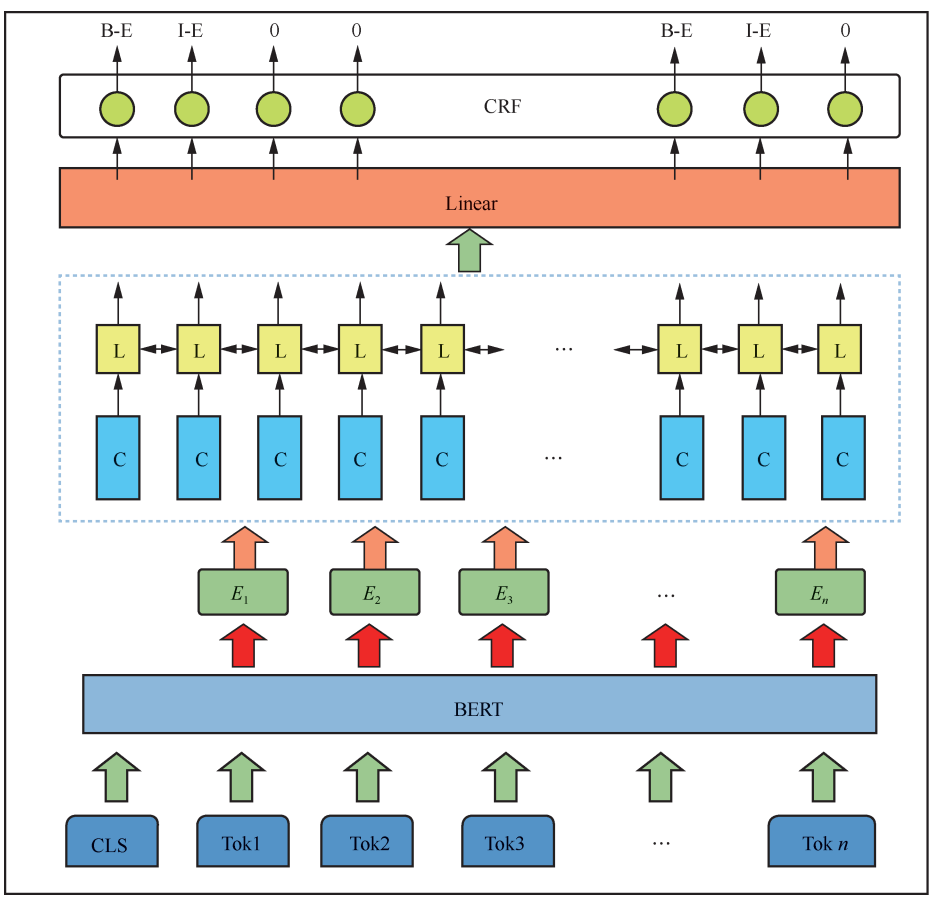

图 5 享受主体识别模型结构

$$\boldsymbol{X}_{\text{hidden}} = \text{BERT}(\boldsymbol{S}) \quad (20)$$

步骤4：使用BiLSTM将隐藏层参数降维为k，接着使用全连接层将维度降至标签个数t。

$$\boldsymbol{X}_{\text{hidden}} = \text{BiLSTM}_k(\boldsymbol{X}_{\text{hidden}}) \quad (21)$$

$$\boldsymbol{Y} = \text{Linear}_t(\boldsymbol{X}_{\text{hidden}}) \quad (22)$$

步骤5：使用条件随机场（conditional random fields，CRF）对输出层的标签进行约束，输出最优的标注序列[17-18]。

$$\boldsymbol{Y} = \text{CRF}(\boldsymbol{Y}) \quad (23)$$

2.3.2 税种、减免类型、减免方式、政策类型识别

在标注数据的过程中，笔者发现28.3%的条款涉及多个税种，如“五、对青藏铁路公司及其所属单位自用的房产、土地免征房产税、城镇土地使用税。”其中涉及房产税和城镇土地使用税两个税种。因此笔者采用了多标签识别的方法。给定训练集：$\{\boldsymbol{S},\boldsymbol{Y}\}$、词表$V$、标签空间$\boldsymbol{L}=\{l_1,l_2,l_3,\cdots,l_n\}$，第$i$个条款文本表示为$S_i=\{w_1,w_2,w_3,\cdots,w_m\}$，其中$\forall w\in V$。词表$V$除样本集包含的字符外，还包含MASK、CLS、PAD、UNK、SEQ这些无实际语义的特殊字符。$Y_i=\{y_1,y_2,y_3,\cdots,y_n\}$是由0或1组成的列表，$y_i$为1时对应第$i$个标签的税种，标签空间$L=\{l_1,l_2,l_3,\cdots,l_n\}$对应一个映射函数$\chi$：$x\to\hat{L}$，其中$\hat{L}\subseteq L$且$\left|\hat{L}\right|\geqslant 1$。

步骤1：把输入的条款转换为字符级别的序列。

步骤2：序列前端增加CLS符号，生成序列的向量表示。

$$\boldsymbol{S}=\left[\text{cls},t_1,t_2,\cdots,t_{\text{maxlen}-1}\right] \quad (24)$$

步骤3：使用BERT模型对序列嵌入进行特征提取。

$$\boldsymbol{X}_{\text{hidden}} = \text{BERT}(S) \quad (25)$$

步骤4：使用全连接层对隐藏层第1个位置（CLS对应的特征向量）进行特征提取。此向量包括整句的所有语义信息，全连接层将CLS特征向量维度降至标签个数t。

$$\boldsymbol{Y} = \text{Linear}_t\left(\boldsymbol{X}_{\text{hidden}}[0]\right) \quad (26)$$

步骤5：最后使用sigmoid分类器计算相应的标签。

$$\boldsymbol{Y} = \text{sigmoid}(\boldsymbol{Y}) \quad (27)$$

减免类型、减免方式、政策类型均属于文本分类任务，采用与税收优惠主题识别同样的方法进行处理。

2.3.3 有效期限、标题、文号识别

对于有效期限、标题、文号这3个相对简单、规律性强、特征比较突出的关键要素，采用基于规则的算法进行抽取识别。绝大多数税收优惠政策是由国家税务主管部门进行撰写和公布的，因此税收优惠政策的结构和格式有很强的规律性和统一性，见**表2**。经过大量的税收优惠政策总结，笔者共发现有效期限、标题、文号的特征30余种。根据总结特征，分别使用正则算法进行规则匹配，可以有效地提取和识别有效期限、标题和文号3个关键数据字段。

如**图6**所示，在处理完税收优惠主题识别和税收优惠分面识别后，将结果全部输出到税收优惠关键要素数据模板，该模板界面支持识别结果的全览和修正工作，并使用SQL数据库存储和管理数据。

为了方便扩充更多的数据集，本文在设计数据库时结合了数据标注时的场景，分别设计了{BIG_TAX，CLAUSE，CLAUSE_TAX，CLAUSE_ENJOY，

表 2 特征规则示例

字段名称	举例（正则）
有效期限	[0-9]{1,4}年[0-9]{1,2}月[0-9]{1,2}日(?=起执行)、(?<=执行时间为)[0-9]{1,4}年[0-9]{1,2}月[0-9]{1,2}日至[0-9]{1,4}年[0-9]{1,2}月[0-9]{1,2}日……
标题	(?<=\s)[\u4e00-\u9fa5]+[0-9]*[\u4e00-\u9fa5]+(?<=通知|公告)、(?<=\s)[\u4e00-\u9fa5]+(?<=通知|公告)……
文号	(?<=通知|公告)[\u4e00-\u9fa5]{1,10}〔[0-9]{4,}〕[0-9]+号、[\u4e00-\u9fa5]{1,10}〔[0-9]{4,}〕[0-9]+号……

图 6 税收政策处理后结果预览

ENJOY, NOTICE, SMALL_TAX}数据表。在使用者提交经过调整的正确数据后，这些数据表不仅存储了数据信息，同时存储了每个条款对应的数据标签，如“享受主体”字段在条款中的索引位置以及BIO标签、“税种”字段的标签类型等。该系统处理新的税收优惠文档后，数据集也会不断扩充，可以在数据库中导出扩充后的新数据集对模型进行再次训练，在大量、高质量数据集的支持下，该系统的算法模型性能也会进一步提高[19]。

2.4 税收优惠政策查询与可视化

如**图7**所示，税收政策优惠查询与可视化是一种基于结构化数据的应用，使用户能够快速检索与享受主体相关的税收优惠信息，并采用径向布局的可视化方式来展示以纳税人为核心的相关内容。其中，绿色节点代表输入的享受主体，黄色节点代表该享受主体所能享受的税种，橘红色节点代表对应税种纳税人能享受的优惠条款。右侧部分是每个条款关键要素的详情信息，用户可以从中快速了解该条款描述的重要内容。

税收政策优惠查询与可视化的主要过程是：①客户端用户输入待了解的享受主体内容，发送至服务端；②服务端在数据库中匹配享受主体内容，如果没有匹配到当前输入的享受主体，则匹配与该享受主体语义相似的其他享受主体内容并返回客户端（例如，“老师”和“教师”在语义上比较相似，当匹配“老师”失败时，则返回“教师”的信息）；③客户端收到相关内容后，以享受主体为中心进行径向图布局。

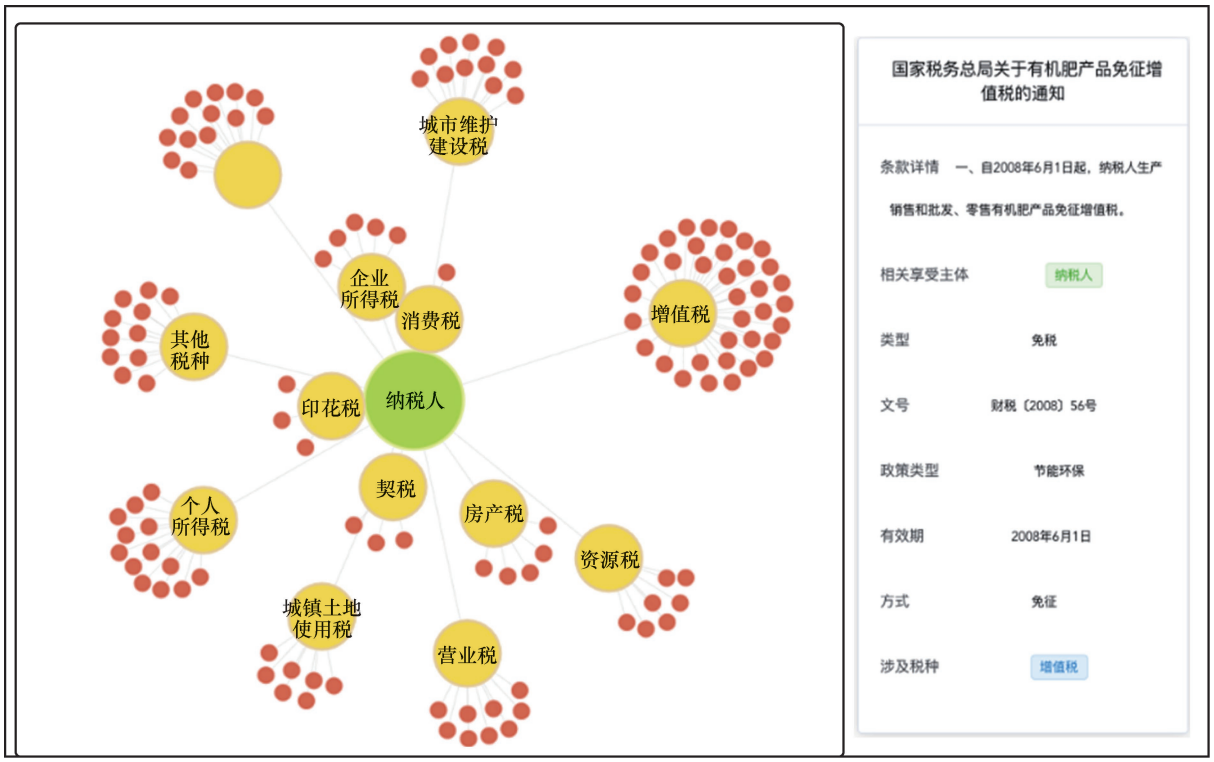

图 7 税收优惠政策可视化查询

享受主体相似匹配算法将莱文斯坦距离（Levenshtein distance）算法作为技术基础。莱文斯坦距离是一种编辑距离算法，通过求出编辑距离，计算两个字符串的相似度 $\text{Similarity}=(\text{Max}(x,y)-\text{Levenshtein})/\text{Max}(x,y)$，其中$x$、$y$为源串和目标串的长度。最后，根据设定的相似度阈值，遍历搜索数据库中所有享受主体，当相似度大于或等于该阈值时，则认为其是当前要查询的享受主体的相似享受主体。

在可视化的实现过程中，大多数可视化工具的内置基础布局并不能完全满足享受主体相关径向图，因此本文采用G6[1]提供的自定义布局算法接口辅助可视化算法的实现。

主体相关径向图 $G=(V,E)$，节点类型为 $V\in V^{\text{e}},V^{\text{t}},V^{\text{q}}$，节点数目为 $n^{\text{e}}=|V^{\text{e}}|$，$n^{\text{t}}=|V^{\text{t}}|,n_i^{\text{t}}=|V_i^{\text{t}}|,i\in 1,\cdots,n^{\text{t}},V^{\text{q}}\in V_i^{\text{t}}$。节点有3种类型：主体节点、税种节点、条款节点。一个主体对应多个税种，相应的一个税种对应多个条款。在主体相关径向图中，以一种享受优惠政策的享受主体为焦点并将其布局在图的中心，相关税种距离为一度，各税种相关的条款距离为二度进行布局。

步骤1：可视化布局，以享受主体为中心，享受主体圆心半径如下。

$$P_{\text{c}}:(x^{\text{c}},y^{\text{c}}),d^{\text{c}}=R_1 \qquad (28)$$

步骤2：计算一度布局，一度布局描述的是与享受主体相关的税种，其围绕在享受主体外一层附近的环上。

首先，计算单位偏移角度k，然后根据偏移角度k，按照顺序依次计算各个节点的

[1] G6是一个简单、易用的图可视化引擎，它提供了图的绘制、布局、分析、交互、动画等图可视化的基础功能，相比于其他可视化工具，G6在关系图形方面具有更多的类别选择和更强的可操作性。

坐标位置。因为不同的税种对应的条款数目不同，所以其与享受主体的距离不一样，距离与 n_i^{t} 有关，即条款数目越多，距离圆心越远。α 为调整距离比的参数，享受主体与任一税种节点的直径之和不大于包含最大条款数目税种与参数 α 的乘积。

$$k=\frac{2\pi}{n^{\mathrm{t}}} \tag{29}$$

$$x_i^{\mathrm{t}}=\mathrm{x}^{\mathrm{c}}+\alpha n_i^{\mathrm{t}}\cos(i\times k) \tag{30}$$

$$y_i^{\mathrm{t}}=\mathrm{y}^{\mathrm{c}}+\alpha n_i^{\mathrm{t}}\sin(i\times k) \tag{31}$$

满足：

$$d^{e}+d^{\mathrm{t}}<\alpha\max(n_i^{\mathrm{t}}) \tag{32}$$

圆心直径如下：

$$p_i^{\mathrm{t}}:\left(x_i^{\mathrm{t}},y_i^{\mathrm{t}}\right) \tag{33}$$

$$d^{\mathrm{t}}=R_2 \tag{34}$$

步骤3：计算二度布局，二度布局是指每个税种节点对应的条款节点围绕在与之对应的税种节点外层的环状布局。

$$u=\frac{2\pi}{n^{\mathrm{t}}} \tag{35}$$

$$x_j^{\mathrm{q}}=x_i^{\mathrm{t}}+\cos\left(\frac{u}{2}\times j\right) \tag{36}$$

$$y_j^{\mathrm{q}}=y_i^{\mathrm{t}}+\sin\left(\frac{3u}{2}\times j\right) \tag{37}$$

$$j\in 1,\cdots,n_i^{\mathrm{t}} \tag{38}$$

$$P_j^{\mathrm{q}}:\left(x_j^{\mathrm{q}},y_j^{\mathrm{q}}\right) \tag{39}$$

3 税收优惠政策数据集

3.1 数据集说明

系统开发阶段用到了许多数据集，具体说明如下。

（1）税收优惠政策法规数据集

数据来源于国家税务总局网站以及各省市地方税务分局官方网站等，包括1990—2020年发布的税务优惠政策4 000余篇文档。每个文档平均包含996个汉字，经过文本分割算法切分条款共计12 000余条。为了给系统提供减免税主题识别的功能，笔者根据需要筛选并标注了2 000条数据用于训练。

（2）享受主体识别数据集

该数据集对识别享受税收优惠政策的纳税人提供数据支撑。目前专业领域的中文数据集尚为稀缺，因此笔者针对税务领域纳税实体标注了2 000余条包含税收优惠的减免税主题条款。

（3）税种多标签分类数据集

该数据集为识别税收优惠政策涉及的税种提供数据支撑。对于该数据集的构建，笔者通过统计4 000余篇税收优惠政策文档包含的税种类型，同时结合税务主管部门官方提供的税种分类体系，在数据集构建过程中，共设立并标注税种标签19种，其中包含：增值税、消费税、企业所得税、个人所得税、资源税、城市维护建设税、房产税、印花税、城镇土地使用税、土地增值税、车船税、车辆购置税、烟叶税、耕地占用税、契税、环境保护税、进出口税收、营业税、其他税种。

（4）其他税收优惠条款分面识别数据集

这部分数据集与上述数据集类似，只是在上述数据集原有的基础上做了更多的分类标注和实体标注。

3.2 实验对比结果

本文实验是基于第3.1节的数据集开展的。笔者使用不同的方法对比任务类型相同的关键要素。本文将精确率（precision）、召回率（recall）以及F1分数（F1 score）作为评估指标。

$\text{precision}=\frac{TP}{TP+FP}$，$\text{recall}=\frac{TP}{TP+FN}$，$F1=2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$，其中TP、FP、FN分别表示真阳率、假阳率、假阴率。

实验设置：深度学习实验框架为PyTorch 1.10 Release，预训练语言模型BERT为Bert-Base-Chinese版本，词表大小为 21 128个词，隐藏层数为12，词嵌入向量维度为768，注意力机制为12个。将数据集中的数据顺序随机打乱，将其中80%作为训练集，剩余20%作为测试集。训练时采用十折交叉验证，将训练数据集分成10组，每次使用9组训练模型，1组进行验证，一共进行10次训练，最后取10次验证的平均值作为最终的分数。其中，文本分类任务、多标签任务和序列标注任务的参数设置如下：学习率为0.0001，批次大小为16，迭代次数为50，可处理序列最大长度maxlen为512；BiLSTM模块的参数设置如下：隐藏输出维度为256，隐藏层数为1，丢弃率为0.3；Linear模块的参数设如下：输出维度为2；BiGRU模块的参数设置如下：隐藏输出维度为256，隐藏层数为1，丢失率为0.3；CRF的参数设置如下：标签数为2。

标题、文号、有效期限的抽取属于基于规则的任务。该任务笔者把抽取内容与原目标内容进行比较，如果相同，则标记为1，否则标记为0。从数据中随机抽样10组，每组为总数据的20%，将precision作为评价指标，实验结果见**表3**。

条款内容、减免类型、减免方式、政策类型属于文本分类任务。该实验选择了两种文本分类的方法Fasttext和TextCNN进行比较。评价指标采用F1分数、precision、recall。见**表4**，在其他处理方法相同的情况下，BERT模型的处理结果优于Fasttext和TextCNN的处理结果。

税种的识别属于文本多标签分类任务。该任务将汉明损失（Hamming loss）作为指标。

$$\text{Hamming loss}=\frac{1}{|\Gamma|}\sum_{i}^{|\Gamma|}\frac{\text{xor}(Y_{ti},Y_{pi})}{|L|} \quad (40)$$

式（40）的结果表示所有标签中错误样本的比例，该值越小，则分类器的分类能力越强。其中$|L|$表示标签总数，$|\Gamma|$表示样本总数，xor表示异或运算。如**图8**所示，在处理数据时笔者发现，各税种数量的高度不均衡导致了长尾效应。因此采用分步处理的方式，首先使用分类模型判断条款中的税种数目，如果该数目大于4个，则使用基于规则的方法进行识别，否则使用文本多标签

表3 规则抽取结果

关键要素	标题	文号	有效期限
precision	0.939（avg）	0.896（avg）	0.933（avg）

表4 分类结果对比

关键要素	BERT			Fasttext			TextCNN		
	F1分数	precision	recall	F1分数	precision	recall	F1分数	precision	recall
条款内容	0.914	0.921	0.898	0.505	0.923	0.347	0.825	0.83	0.825
减免类型/方式	0.97	0.99	0.96	0.857	0.897	0.82	0.663	0.964	0.611
政策类型	0.831（Micro）	0.83	0.847	0.864（Micro）	0.592	0.646	0.675（Micro）	0.658	0.711

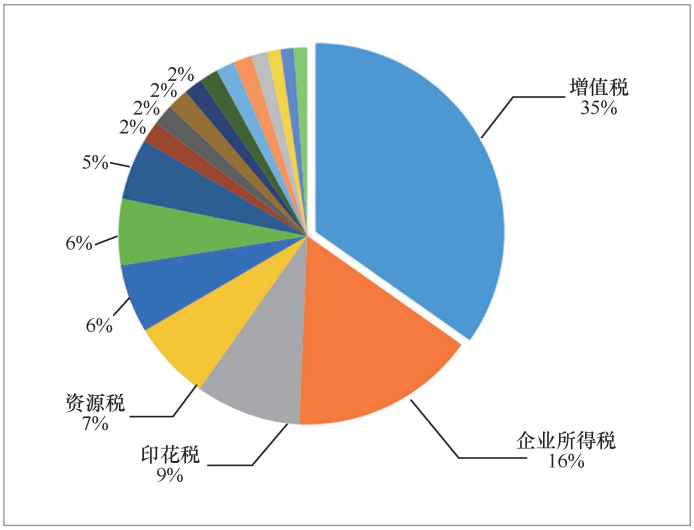

图 8 各个税种比例分布情况

分类方法进行识别。实验结果（见**表5**）表明，相比于直接使用BERT模型的方法，该处理方式的效果有所提升。

享受主体的抽取属于序列标注任务，该实验分别采用BERT+BiLSTM+CRF、BERT+Linear+CRF和BERT+BiGRU+CRF 3种不同的方法进行对比，具体实验结果见**表6**。

4 结束语

本文设计开发了一个使用简单、操作便捷的税收优惠法规可视化分析系统。该系统设计了税收优惠关键要素数据模板，定义了税收优惠主题和税收优惠分面，实现了主题和分面的识别和抽取工作，完成了由非结构化的税收优惠政策到结构化的税收优惠关键要素数据模板的转换，研究开发了税收优惠法规查询与可视化分析功能。纳税人可以通过该系统查询所有与自身利益相关的税收优惠政策。并且，该系统的信息抽取功能可以经过人工干预，完成数据的修正调优，存储后的数据可以作为该系统的新数据集再次训练，从而使系统的识别与处理精度继续提升。

目前的工作只是一个开始，下一步笔者计划在税务领域做出更多的工作。首先是不断地扩大税务领域的数据集，同时在方法层面继续改进，进一步提高模型性能；其次，笔者的目标是把当前的工作应用到税务常识图谱的构建中，通过构建该图谱能够更好地为税务智能查询以及税务智能问答等任务提供有力支撑。

参考文献:

[1] 汪疆平，肖戎. 税务大数据分析的技术和典型应用[J]. 大数据，2017，3(2)：92-103.
WANG J P，XIAO R. Big data analysis technology and application on taxation[J]. Big Data Research，2017，

表 5 多标签文本分类结果

关键要素	模型直接处理的Hamming Loss	分步处理后的Hamming Loss
税种	0.023	0.016

表 6 序列标注结果对比

关键要素	BERT+BiLSTM+CRF（本文方法）			BERT+Linear+CRF			BERT+BiGRU+CRF		
	F1分数	precision	recall	F1分数	precision	recall	F1分数	precision	recall
享受主体	0.73	0.71	0.75	0.69	0.67	0.71	0.71	0.68	0.75

3(2): 92–103.

[2] 邵凌云. 基于纳税人需求 优化纳税服务机制[J]. 税务研究, 2013(5): 76–79.
SHAO L Y. Optimize the tax service mechanism based on the demand of taxpayers[J]. Taxation Research, 2013(5): 76–79.

[3] 谢学刚, 苟仁金. 提升纳税服务质量的现实选择[J]. 税务研究, 2014(11): 96.
XIE X G, GOU R J. A realistic choice to improve the quality of tax services[J]. Taxation Research, 2014(11): 96.

[4] 谢波峰. 基于大数据的税收经济分析和预测探索[J]. 大数据, 2017, 3(3): 15–24.
XIE B F. Exploratory research on big data application of analysis and forecasting in economics of tax[J]. Big Data Research, 2017, 3(3): 15–24.

[5] ZHANG R X, YANG W, LIN L Y, et al. Rapid adaptation of bert for information extraction on domain-specific business documents[J]. arXiv preprint, 2020, arXiv:2002.01861.

[6] NGUYEN M T, LE D T, LINH L T, et al. AURORA: an information extraction system of domain-specific business documents with limited data[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2020: 3437–3440.

[7] FRIEDRICH A, ADEL H, TOMAZIC F, et al. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1255–1268.

[8] ZEGHDAOUI M W, BOUSSAID O, BENTAYEB F, et al. Medical-based text classification using FastText features and CNN-LSTM model[C]// Database and Expert Systems Applications. Cham: Springer, 2021: 155–167.

[9] BRANDES U, KENIS P, WAGNER D. Communicating centrality in policy network drawings[J]. IEEE Transactions on Visualization and Computer Graphics, 2003, 9(2): 241–253.

[10] BRANDES U, PICH C. More flexible radial layout[J]. Journal of Graph Algorithms and Applications, 2011, 15(1): 157–173.

[11] RAJ M, WHITAKER R T. Anisotropic radial layout for visualizing centrality and structure in graphs[C]//Graph Drawing and Network Visualization. [S.l.:s.n.], 2018: 351–364.

[12] FENU G, SPANO L D. Recommendation Centre: inspecting and controlling recommendations with radial layouts[C]//Workshop on Engineering Computer-Human Interaction in Recommender Systems. [S.l.:s.n.], 2016: 54–61.

[13] BOSTOCK M, HEER J. Protovis: a graphical toolkit for visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1121–1128.

[14] BOSTOCK M, OGIEVETSKY V, HEER J. D^3 data-driven documents[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2301–2309.

[15] LI D Q, MEI H H, SHEN Y, et al. ECharts: a declarative framework for rapid construction of web-based visualization[J]. Visual Informatics, 2018, 2(2): 136–146.

[16] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of NAACL-HLT, 2019: 4171–4186.

[17] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. Morristown: Association for Computational Linguistics, 2003: 188-191.

[18] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint, 2015, arXiv:1508.01991.

[19] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: a survey[J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.

作者简介

关海山(1996-),男,西安交通大学软件学院硕士生,主要研究方向为自然语言处理、文本问题生成。

郑玉龙(1996-),男,西安交通大学软件学院硕士生,主要研究方向为自然语言处理、文本问题生成。

魏笔凡(1977-),男,博士,西安交通大学继续教育学院研究员,主要研究方向为Web信息抽取、教育知识图谱构建及应用。

张泽民(1999-),男,西安交通大学软件学院硕士生,主要研究方向为自然语言处理、深度学习、问题生成等。

岳浩（1999- ），男，西安交通大学软件学院硕士生，主要研究方向为大数据、深度学习等。

师斌（1992- ），男，博士，西安交通大学计算机科学与技术学院讲师，主要研究方向为金融数据挖掘、云计算及虚拟化技术。

董博（1983- ），男，博士，西安交通大学计算机科学与技术学院高级工程师，主要研究方向为金融数据挖掘、智能教育。

收稿日期： 2022-11-15

通信作者： 魏笔凡，weibifan@xjtu.edu.cn

基金项目： 国家重点研发计划资助项目（No.2020AAA0108800）；国家自然科学基金资助项目（No.62137002，No.61937001，No.62176209，No.62176207，No.62106190，No.62050194）；国家自然科学基金创新研究群体资助项目（No.61721002）；教育部创新团队资助项目（No.IRT_17R86）；中国工程院咨询研究资助项目“基于MOOC中国的‘一带一路’人才培养的线上线下混合教学支撑信息化平台与服务体系”；中国博士后科学基金项目（No.2020M683493）；中国工程科技知识中心资助项目

Foundation Items: The National Key Research and Development Program of China (No.2020AAA0108800), The National Natural Science Foundation of China (No.62137002, No.61937001, No.62176209, No.62176207, No.62106190, No.62050194), Innovative Research Group of the National Natural Science Foundation of China (No.61721002), Innovation Research Team of Ministry of Education (No.IRT_17R86), Consulting Research Project of Chinese Academy of Engineering “The Online and Offline Mixed Educational Service System for ‘The Belt and Road’ Training in MOOC China”, China Postdoctoral Science Foundation (No.2020M683493), Project of China Knowledge Centre for Engineering Science and Technology