基于异构图神经网络的检务知识咨询业务分类

蔡惠民[1,2*],印忠文[1,2],岳世彬[1,2]
(1. 中电科大数据研究院有限公司,贵州 贵阳 550022;
2. 提升政府治理能力大数据应用技术国家工程实验室,贵州 贵阳 550022)

摘 要:随着检察机关办理的案件日益增多,用户对检务领域知识的咨询需求逐年增大,传统依靠检察机关领域专家人工解答回复的方式难以应对大规模的咨询服务。为了提高用户咨询服务效率,提升计算机正确理解用户提问意图的能力,提出了一种面向检务知识咨询的异构图神经网络业务类型分类模型。该模型以基于句法依存分析的图表示、基于邻域窗口的图表示作为输入,分别以 RGCN 和 GAT 图神经网络作为特征编码器,并通过特征融合实现用户提问内容业务类型预测。同时引入辅助分类器优化特征编码器的学习并提升模型性能,并采用 Focal Loss 损失函数有效解决了样本数据不均衡问题。此外,该模型与传统深度学习文本分类模型、目前主流的 BERT 模型在宏平均准确率、模型大小、推理时间等多个维度进行性能对比。对比实验显示,该模型在测试集上的宏平均准确率均优于文本分类基准模型,模型大小和推理时间远小于 BERT 模型。

关键词:智慧检务;异构图神经网络;意图识别;业务分类;检答网
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)10-0201-08
doi:10.3969/j.issn.1673-629X.2022.10.033

Heterogeneous Graph Neural Network Based Business Classification for Procuratorial-knowledge Query

CAI Hui-min[1,2*], YIN Zhong-wen[1,2], YUE Shi-bin[1,2]
(1. CETC Big Data Research Institute Co., Ltd., Guiyang 550022, China;
2. Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, Guiyang 550022, China)

Abstract: With the increasing of cases handled by the procuratorate, the user's demand for consulting knowledge in the field of procuratorial affairs is raised year by year. However, it is difficult to deal with the large-scale consulting services by relying on the experts in procuratorate to answer and reply manually. In order to improve the efficiency of user consultation and enhance the ability of computer to correctly understand the user's query, a heterogeneous graph neural network model is proposed for business type classification. This model uses graph representations based on both syntactic dependency analysis and neighbor window as inputs. RGCN and GAT graph neural networks are used as feature encoders respectively, and feature fusion is applied to predict the business type of user's query. Meanwhile, auxiliary classifiers are introduced to optimize the learning of feature encoders and improve the performance of the model. Moreover, Focal Loss is applied to effectively address the issue of unbalanced data. Furthermore, the performance of this model is compared with the traditional deep learning text classification models and the current mainstream BERT model in many dimensions, such as macro average accuracy, model size, reasoning time. Comparative experiments show that the macro average accuracy of the proposed model on the test set is better than other benchmark text classification models. The model size and reasoning time of the proposed model are much smaller than BERT model.

Key words: wisdom procuratorial affairs; heterogeneous graph neural network; intention recognition; business classification; Jianda-Net

0 引 言

2017 年以来,最高人民检察院先后出台了《关于深化智慧检务建设的意见》、《全国检察机关智慧检务行动指南 2018-2020 年》等重要文件,为打造智慧检务明确了发展方向。智慧检务是指运用大数据、人工智能等新兴技术,通过对司法数据的有机整合与智能

收稿日期:2021-11-19 **修回日期**:2022-03-22
基金项目:国家重点研发计划项目(2020YFC0833003);天津市科技计划项目(19YFZCGX00680)
作者简介:蔡惠民(1985-),男,博士,CCF 会员(98979M),通信作者,研究方向为自然语言处理、知识图谱及其应用等。

分析,挖掘数据的潜在价值,使其服务于司法应用,推动更高形式的检察信息化建设,对于辅助科学决策、提升办案效能、规范司法办案、推进司法改革等有重要意义[1-2]。

随着大数据时代的到来,检务大数据的积累为进一步推进智慧检务建设奠定了坚实基础。检答网是检察人员内部业务研讨交流平台。其作为检务信息化建设的重要组成部分,多年来积累了大量用户对检察业务知识的咨询数据,以及检察机关各级领域专家对用户问题的解答与回复数据。然而,随着检察机关办理的案件日益增多,基层检察办案人员等用户对检务知识咨询需求不断增大,仅仅依赖检务领域专家对其人工回答需要投入大量的人力成本。其次,对问题的回复往往需要遵循特定的流程规范和内容审核,使用户的问题得不到及时解决。同时,很多常规检务知识的提问频次较高,存在重复性人工解答的现象。

为了提高用户咨询服务效率,增强计算机正确理解用户提问意图的能力,并准确预测用户提问内容所属的业务类型,是构建智能问答系统的关键环节。因此,该文将基于检答网用户提问数据,提出一种面向检务领域用户咨询的业务类型分类模型。首先对检答网原始数据中的业务类型重新进行梳理与归并,构建数据集。其次,基于句法依存分析得到用户提问内容的图表示,并应用 RGCN 图神经网络模型[3]提取其特征。同时基于邻域窗口得到用户提问内容的图表示,将 GAT 图神经网络模型[4]作为其特征编码器。最后,构建一种融合两种图表示特征的异构图神经网络模型,并通过引入辅助分类器优化模型性能,采用 Focal Loss 损失函数[5]解决样本数据的不均衡问题,实现对用户提问内容的业务类型预测与性能评估,为进一步构建面向检务领域智能问答系统打下坚实的基础。

1 相关工作

文本分类是自然语言处理的基础问题。与传统基于朴素贝叶斯方法[6]、支撑向量机(SVM)[7-8]等文本分类方法相比,以卷积神经网络(CNN)[9-12]、循环神经网络(RNN)[13-14]、长短期记忆神经网络(LSTM)[15-16]等为代表的深度学习模型提供了一种端对端的文本分类方法,以数据驱动的方式自动学习文本中潜在的语义模式,避免了人工构建特征的繁琐工作,并获得了更优性能。自 2018 年提出 BERT 模型[17]以来,以 BERT 模型为基础的多种自然语言处理任务均获得较大性能提升。文献[18]利用预训练 BERT 模型提取文本的字符特征,作为文本分类器的输入。文献[19]提出了多种基于 BERT 模型的微调方法,使其应用于文本分类。文献[20]则将 BERT 模型用于中文短文本分类。

近年来以图卷积神经网络为代表的图神经网络模型得到了关注和发展[21-22]。图神经网络模型不仅保留了传统卷积神经网络的优良特性,同时具有能适应图数据的特点,使深度学习技术与图数据的有效结合成为必然。图神经网络通过迭代聚合邻域节点特征而学习到图数据中各节点的特征向量,从而支撑节点分类任务和图分类任务。图神经网络模型与自然语言处理技术的结合也成为一种趋势。文献[23]提出了基于词与词之间的互信息,以及词与文档之间的 TF-IDF 权重构建整个文本语料库图网络,并通过图神经网络模型实现对语料库图网络文档节点的分类。然而,这种构建大规模文本图网络实现节点分类的方式虽然能利用语料库中全局信息,但并不适合模型的在线部署,同时存在较大的内存消耗。为此,文献[24]通过词的邻域窗口构建文档的图表示,并提出了基于文档的图神经网络分类模型。文本的句法依存关系也用于文本的图表示。文献[25]通过将图神经网络用于句法依存图,实现机器翻译。文献[26-27]则将图神经网络与句法依存树相结合,用于事件抽取任务。

智能问答系统中的意图识别通常需要解决用户提问文本内容的领域分类[28]。针对检答网用户提问内容的长度短等特点,多样化的图表示有利于充分挖掘短文本的有用信息,因此尝试构建能融合不同图表示的异构图神经网络,以用于面向检务知识咨询的文本分类任务。

2 模型与方法

2.1 基于句法依存分析的图表示

针对检答网中的用户提问数据,用户提问内容的长度不一:有些提问简短,只包含一句话;有些提问的描述较为具体,可能包含多句话。因此,本节先从句子粒度考虑,通过句法依存分析构建单个句子的图表示。针对多句话的用户问题,通过单句的图表示构建多句的图表示。

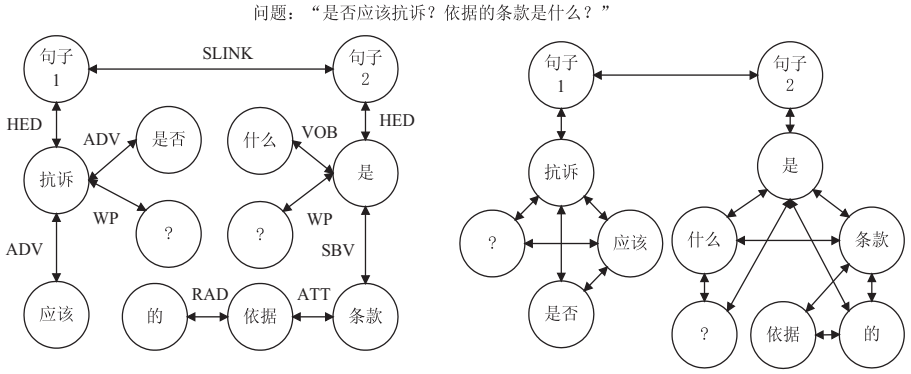
2.1.1 单句的图表示

针对中文文本,当前句法依存树的提取技术较为成熟,该文采用哈工大语言技术平台的 LTP 工具,用于检答网中用户提问内容的句法依存分析,并基于提取的句法依存树构建单句的图表示。具体为:首先以句子为单位,将用户提问内容切分为多个句子的集合。针对每个句子,应用 LTP 工具得到分词后词与词之间的句法依存关系及其指向关系。以分词后每个词为图节点,节点的特征初始化为 Word2vec 预训练模型[29]的词向量。基于句法依存指向关系列表,连接词与词之间存在指向关系的所有边,句法依存关系即定义为

边的类型。考虑到基于句法依存分析的边稀疏性,为了利于图神经网络的特征聚合,将原有句法依存指向性的单向边更改为双向边。同时添加一个句子节点,该节点连接句法依存关系为“HED”的词汇节点。此外,句子节点的特征向量初始化为 Word2vec 预训练模型中“起点”的词向量。通过这种方式,构建单句的图表示。

2.1.2　多句的图表示

针对多句话的用户问题,假设已通过句法依存分析工具得到单句的图表示,该文通过双向连接相邻句子的句子节点,从而构建多句的图表示。其中该双向连接的边类型定义为“SLINK”,如图 1 所示。图 1 的 A 部分示意了用户提问“是否应该抗诉? 依据的条款是什么?”经过句法依存分析后的多句图表示。该图表示的边均为双向边,边的类型取决于句法依存分析的结果,句子与句子之间通过类型为“SLINK”的双向边连接。



(A) 基于句法依存的图表示　　(B) 基于邻域窗口的图表示

图 1　基于句法依存分析及基于邻域窗口的多句图表示($d=2$)

2.2　基于邻域窗口的图表示

基于句法依存分析的图表示虽然将用户提问内容分解为词与词之间的句法关系,但却忽略了词与词之间的邻域关系。而词的邻域关系在词向量学习中有广泛应用。同时,HUANG 等提出了将词的邻域关系应用于图神经网络,证明了其可行性与有效性。因此,该文同时考虑基于邻域窗口的图表示。

2.2.1　单句的图表示

基于邻域窗口的图表示方式仍以分词后每个词为图节点,节点的特征初始化为 Word2vec 预训练模型的词向量。假设词的邻域定义为与词的距离不大于 d 的词集合,则基于邻域窗口的图表示构建规则为:单句分词后每个词为图的节点,每个词与其邻域中每个词建立双向边连接,从而得到该单句的图表示。考虑到用户提问内容长度较短,该文将 d 设为 2。

2.2.2　多句的图表示

针对检答网中用户提问数据的多句情况,仍采用 2.1 节中多句图表示的策略。即通过双向边连接相邻句子的句子节点,使句子之间的信息可以交流。对于每个句子,句子节点与句法关系为“HED”的词节点相连,与 2.1 节保持一致,如图 1 所示。图 1 的 B 部分示意了用户提问内容基于邻域窗口的多句图表示。该图表示的边均为双向边,句子与句子之间也通过双向边连接,但不考虑边类型。

2.3　基于异构图的图神经网络模型

该文同时考虑基于句法依存分析的图表示,以及基于邻域窗口的图表示。从不同维度提取用户提问内容的结构信息,充分利用短文本的有限信息,有利于对提问意图的正确理解。由于这两种图表示方法的差异性,提出一种基于异构图的图神经网络模型,将提取并融合这两种图表示的特征。

对于基于句法依存分析的图表示,该表示方法保留了边的多种句法依存关系。而 RGCN 编码器对带有多类型边的图有较强的建模能力,其为不同的关系类型引入不同的权重参数,能充分利用边的类型信息增强图神经网络对特征的编码能力。如公式(1)所示,对于边类型为 r 的情况,RGCN 编码器对第 l 层中所有边类型为 r 的邻域节点特征 h_j^l 进行聚合,最后通过非线性变换 σ 得到第 $l+1$ 层的节点特征 h_i^{l+1}。$|N^r(i)|$ 表示节点 i 中边类型为 r 的邻域大小,用于归一化特征聚合。

$$h_i^{l+1}=\sigma\left(\sum_{r\in R}\sum_{j\in N^r(i)}\frac{1}{|N^r(i)|}W_r^l h_j^l+W_0^l h_i^l\right) \tag{1}$$

该文应用两层 RGCN 编码器对基于句法依存分析的图表示提取特征,如图 2 中 M1 部分所示。由于用户提问内容的长度较短,而基于句法依存分析的所构建的图较简单,不采用更多的 RGCN 层有利于避免图神经网络的过平滑效应。每层 RGCN 编码器后都经过 ELU 非线性变换以及 DropOut 层。最后通过 ReadOut 层得到图表示的特征编码。其中 ReadOut 层定义为输入全局最大池化和输入全局平均池化的

拼接。

而对于基于邻域窗口的图表示,其中每个节点最多有 $2d$ 条边与邻域节点连接,该文应用多层 GAT 编码器提取其特征。每层 GAT 编码器将学习每个邻域节点的重要性权重。如公式(2)所示,对于第 l 层 GAT 编码器,每条边的权重由 $a_{i,j}^{l}$ 决定。而权重 $a_{i,j}^{l}$ 大小取决于第 l 层节点特征 h_i^l 和邻域节点特征 h_j^l。特征 h_i^l 和特征 h_j^l 分别通过待学习的映射矩阵 $\boldsymbol{W}_0^l$ 变换而拼接后,与一个待学习向量 v_l 取内积,并通过非线性激活函数 LeakyReLU 获得其重要性权重。$a_{i,j}^{l}$ 即为邻域范围内的归一化重要性权重。最后,第 $l+1$ 层节点特征 h_i^{l+1} 根据公式(3)对邻域 $N(i)$ 内所有节点特征进行重要性加权特征聚合,并通过非线性变换 σ 而获得。

$$a_{i,j}^{l}=\frac{\exp(\mathrm{LeakyReLU}(v_l^{\mathrm{T}}[\boldsymbol{W}_0^l h_i^l \parallel \boldsymbol{W}_0^l h_j^l]))}{\sum_{k\in N(i)}\exp(\mathrm{LeakyReLU}(v_l^{\mathrm{T}}[\boldsymbol{W}_0^l h_i^l \parallel \boldsymbol{W}_0^l h_k^l]))} \tag{2}$$

$$h_i^{l+1}=\sigma\Big(\sum_{j\in N(i)} a_{i,j}^{l} W_1^l h_j^l\Big) \tag{3}$$

基于句法依存分析的图表示与基于邻域窗口的图表示分别经过特征编码后,即各自 ReadOut 层的输出向量通过拼接的融合方式作为输出层的输入。输出层由两层全连接层组成。第一层全连接层后经 ELU 非线性变换以及 DropOut 层,而第二层全连接层通过 Softmax 层输出各类的预测概率,如图 2 中 M3 部分所示。

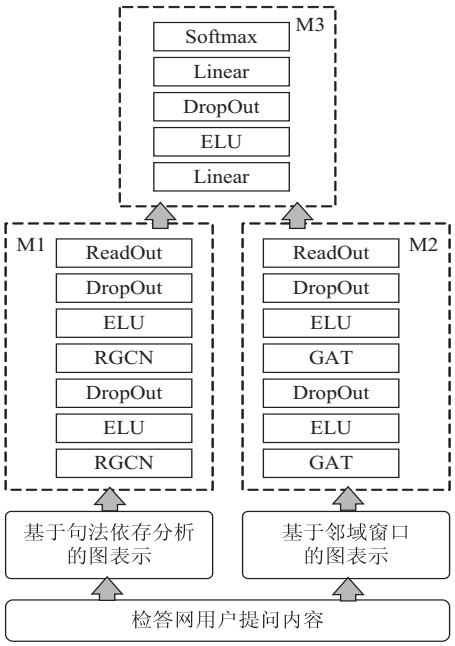


图 2 面向检答网用户咨询问题分类的异构图神经网络模型

2.4 异构图神经网络的训练

为了应对异构图特征融合和训练样本不均衡等问题,该文针对异构图神经网络模型,通过引入辅助分类器[30]来增强底层网络的特征学习能力,有效防止梯度消失;同时通过引入 Focal Loss 损失函数应对训练样本的不均衡问题。

该文分别针对图 2 中 M1 部分的特征编码器和 M2 部分的特征编码器额外添加全连接层和 Softmax 层,作为两个辅助分类器的输出,如图 3 所示。因此,M1 部分的 ReadOut 层引出的辅助分类器对应损失函数 Loss_1,原 M3 部分的分类器对应损失函数 Loss_2,而 M2 部分的 ReadOut 层引出的辅助分类器对应损失函数 Loss_3。总的损失函数 Loss 通过 Loss_1、Loss_2 和 Loss_3 加权求和而得。如公式(4)所示,Loss_1 和 Loss_3 赋予相同的权重 α,则 Loss_2 的权重为 $1-2\alpha$,其中权重 α 位于 0 到 0.5 区间。辅助分类器的引入将增加反向传播的梯度信号,并增强了正则化效果,有利于底层 M1 部分特征、底层 M2 部分特征的学习。

$$\mathrm{Loss}=\alpha\times\mathrm{Loss}_1+(1-2\alpha)\times\mathrm{Loss}_2+\alpha\times\mathrm{Loss}_3 \tag{4}$$

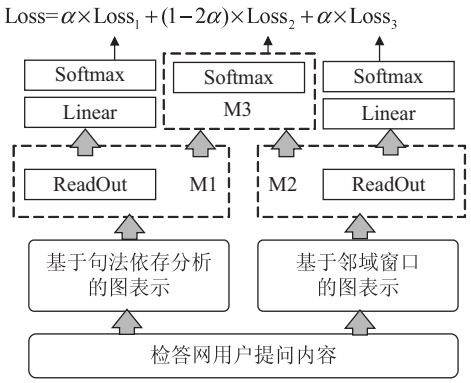


图 3 异构图神经网络的辅助分类器示意图

由 2.1 节可知,检答网用户提问内容归并为 16 个类别,而这些类别存在样本数量不均衡的问题,且有些类别的样本数量差异较大。解决训练样本不均衡的方法有很多,该文主要应用 Focal Loss 损失函数,如公式(5)所示,其中 N 代表总样本数量,K 代表总类别数,$y_{n,k}$ 为第 n 个样本属于类别 k 的真值,$\mathbf{I}$ 为指示函数,$p_{i,n,k}$ 为图 3 中与 Loss_i 对应的 M 模块关于分类类别为 k 的预测概率。通过对样本数量较少的类别 k 赋予较大的权重 β_k 来平衡其反向传播中的梯度信号大小。其中各类别的权重取值策略为:各类别的归一化权重 β_k 正比于自身样本数 N_k 倒数的平方根。取倒数的平方根是为了防止权重差异较大对模型训练带来的不稳定性,如公式(6)所示。另一方面,γ 用于鼓励提高困难样本对梯度的贡献,而减少简单样本的权重,该文 γ 取值为 1。

$$\mathrm{Loss}_i=-\frac{1}{N}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\mathbf{I}\,(y_{n,k}=1)\beta_k(1-p_{i,n,k})^{\gamma}\log(p_{i,n,k}) \tag{5}$$

$$\begin{cases} \sum_{k=0}^{K-1} N_k = N \\ \beta_k = \sqrt{1/N_k} / \sum_{k=0}^{K-1} \sqrt{1/N_k} \end{cases} \tag{6}$$

3　实验与评测

3.1　数据集构建与评估标准

该文以检答网用户提问的文本数据为研究对象,其提问内容覆盖了全国各省市检察院检察办案人员以及基层检察办案人员对检务领域的知识咨询。原始数据共包含了 53 362 条数据,每条数据包含了脱敏后的用户 ID、提问内容、业务分类等字段。针对用户提问内容的业务分类预测需求,从原始数据中提取了提问内容和业务分类两个字段用于构建数据集。

对用户提问内容的业务类型进行数据统计分析时,发现其存在以下问题:其一,业务类型较多,达到 31 个业务类别;其二,各个业务类型下的数据数量分布极不均衡,其中业务类别“普通犯罪检察”的样本数量达到最多的 18 271 条,而业务类别“铁检”和“公诉二”的样本数量仅为 5 条;其三,业务类型分类存在界限模糊、部分类别重复定义的情况,比如业务类型“未检”和“未成年人检察”应为相同类别,又如业务类型“公益诉讼”和“公益诉讼检察”可归并为相同类别。

因此,为了支撑分类算法模型的构建,对原始数据进行预处理,具体处理内容包括:(1)以人工的方式逐条分析用户提问内容和业务分类信息,排除无效数据,并对错误分类的数据进行重新标注;(2)去除业务类别样本量不足的少数数据;(3)基于检务知识背景,制定统一的业务分类标签体系,对类别重复的数据进行合并和类别标签统一,将业务类别数量从 31 个压缩至 16 个。图 4 显示了 16 个业务类型对应的样本数量分布图,其中类型为“司改”的样本数量仅为 31,因此在图 4 中没有得到清晰显示。

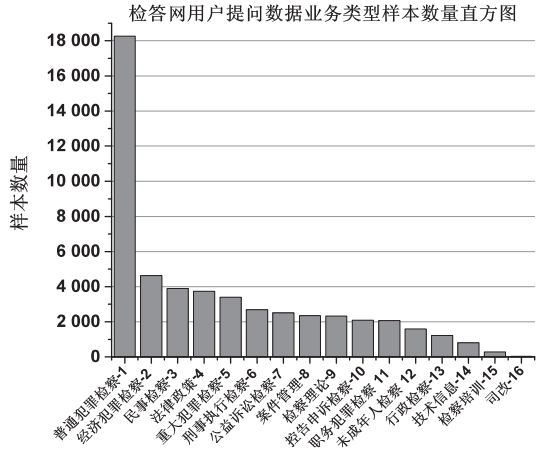


图 4　检答网用户提问数据业务类型样本数量直方图

最后,对预处理后每条数据的顺序随机化,并分别对 16 个业务类别按 7∶1∶2 相同的比例抽取样本形成训练集、验证集和测试集,其大小分别为 36 275、5 196和 10 382。图 5 给出了检答网用户提问内容文本长度直方图。从图可知,检答网用户提问内容以短文本为主。

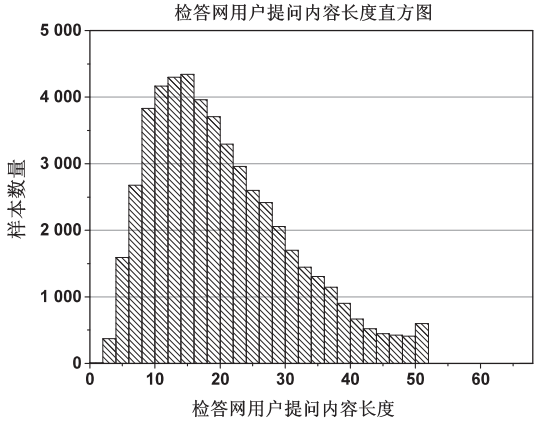


图 5　检答网用户提问内容文本长度直方图

考虑到检答网用户提问内容的业务类别在训练集、验证集和测试集中存在类型样本分布极度不均衡等问题,如图 4 所示,该文将以宏平均-准确率作为算法模型性能的统一评估标准。宏平均-准确率定义为计算各个业务类型分类准确率的平均值。

3.2　模型设置与训练

异构图神经网络参数设置如下:图 3 中的异构图神经网络模型基于 Pytorch Geometric 框架构建。对于 M1 部分的 RGCN 层,针对基于句法依存分析的图表示,边的类型通过遍历训练集中所有用户提问内容经过句法依存分析后,得到的句法依存关系集合,同时添加用于连接句子节点之间的“SLINK”关系类型。考虑到训练集之外可能存在其他的句法依存关系,因此添加一个“Others”关系类型用于应对特殊情况,共 16 个关系类型。为了缩减 RGCN 层的参数,每个 RGCN 层的隐层大小为 64。对于 M2 部分的 GAT 层,每个 GAT 层的隐层大小设置为 64。对于 M3 部分的第一个全连接层,其隐层大小为 128。而第二个全连接层的输出大小为 16,即业务类型的类别数量。

模型训练参数设置如下:异构图神经网络模型的训练采用 Adam 优化器,初始学习率设置为 0.000 1。训练样本的 batch size 设置为 64,epoch 的大小设置为 300。训练过程中,根据模型在验证集上的指标表现确定最终的模型参数。

模型训练与测试的硬件环境为:CPU 型号为 Intel(R) Xeon(R) CPU E5-2620 v4 32 核,内存 64G,GPU 型号为 NVIDIA GTX1080ti。其软件环境为:操作系统为 Ubuntu 16.04.7 LTS,Python 版本为 3.6。

3.3 实验结果

3.3.1 超参数 α 对模型的影响

模型的 RGCN 编码器和 GAT 编码器分别引入辅助分类器,其对应的损失函数为 $Loss_1$ 和 $Loss_3$。为了评估超参数 α 对模型的影响,该文以 0.1 为步长从 0 开始扫描超参数 α ,并记录模型在验证集下的最大宏平均-准确率。如图 6 所示,超参数 α 等于 0 时,模型在验证集上的最大宏平均-准确率为 59.3% ,此时 $Loss_1$ 和 $Loss_3$ 的权重均为 0,等效于不引入辅助分类器的情况。超参数 α 等于 0.3 时,模型在验证集上的最大宏平均-准确率为 61.6% ,高于超参数 α 取其他值时的性能。此时 $Loss_1$ 和 $Loss_3$ 的权重均为 0.3,$Loss_2$ 的权重为 0.4。通过进一步测试,超参数 α 等于 0 和 0.3 时模型在测试集上的宏平均-准确率分别为 58.0% 和 59.3% ,模型性能提升了 1.3% 。该对比实验也说明了通过引入辅助分类器,有助于 RGCN 编码器和 GAT 编码器的特征学习,从而提升了模型的整体性能。因此,模型在后续的性能评估中,超参数 α 固定为 0.3。

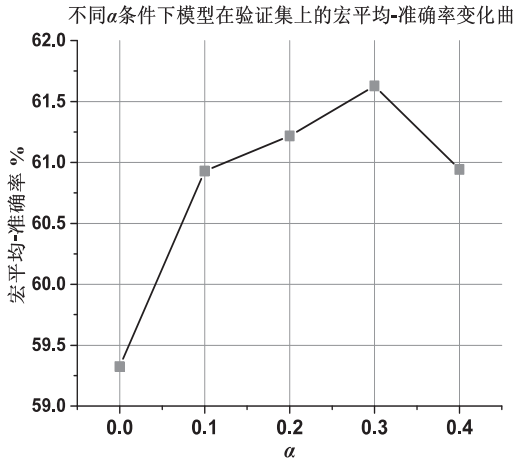


图 6 超参数 α 对模型性能在验证集上的影响

3.3.2 不同编码器对模型的影响

针对检答网用户提问内容,该文采用 RGCN 编码器对基于句法依存的图表示进行特征提取,同时采用 GAT 编码器对基于邻域窗口的图表示进行特征提取,最后通过特征的拼接融合方式实现所属业务类型的分类,该模型定义为 Model-RGCN+GAT。为了评估不同编码器对模型性能的影响,将独立采用 RGCN 编码器的模型定义为 Model-RGCN,模型结构如图 3 中 M1 直接作为 M3 的输入,无辅助分类器和 M2 部分。同时,独立采用 GAT 编码器的模型定义为 Model-GAT,模型结构如图 3 中 M2 直接作为 M3 的输入,无辅助分类器和 M1 部分。三者在测试集上的表现如表 1 所示。从对比结果看,RGCN 编码器提取的特征与 GAT 编码器提取的特征融合后对提升业务类型分类的宏平均-准确率更有帮助。

表 1 不同编码器条件下的模型性能对比

模型	宏平均准确率/%
Model-RGCN	57.4
Model-GAT	57.8
Model-RGCN+GAT	59.3

3.3.3 损失函数对模型的影响

该文引入 Focal Loss 损失函数的目的是为了应对检答网用户提问数据中业务类型样本较大的分布差异。为了评估损失函数对模型性能的影响,如表 2 所示,将 Model-F 定义为引入 Focal Loss 损失函数的情况(与章节 3.3.2 中的 Model-RGCN+GAT 相同),而 Model-C 定义为采用传统交叉熵损失函数的情况。表 2 展示了 Model-F 和 Model-C 在测试集上各个业务类型的分类准确率以及宏平均准确率。由数据对比可知,采用传统交叉熵损失函数时,模型的宏平均准确率为 53.7% 。对于样本数量最少的“司改”类型,模型的预测准确率为 0。而引入 Focal Loss 损失函数后,模型的宏平均准确率为 59.3% ,性能提升了 5.6% 。该模型在各业务类型的预测准确率更均衡,有 13 个类别的预测准确率均高于对照组。该对比实验说明了 Focal Loss 损失函数有效解决了检答网用户提问内容业务类型的样本数量不均衡问题。

表 2 Focal Loss 损失函数与传统交叉熵损失函数对模型的性能对比

类别	1	2	3	4	5	6	7	8	
Model-F	74.1	19.1	70.6	75.8	68.3	75.1	71.1	13.5	宏平均准确率/%
Model-C	88.0	24.4	60.0	78.1	61.0	70.4	68.2	1.9	
类别	9	10	11	12	13	14	15	16	
Model-F	53.0	72.5	58.8	63.8	70.2	66.7	53.6	42.9	59.3
Model-C	46.6	66.6	57.1	58.1	61.8	65.3	51.8	0	53.7

3.3.4 与其他基准模型的性能对比

本节将提出的基于异构图神经网络文本分类模型与传统 CNN、LSTM、Bi-LSTM 文本分类基准模型,同时与近年来主流的 BERT 模型进行性能对比。对比结果如表 3 所示。从对比结果可知,提出的异构图神经网络文本分类模型在测试集上的性能均优于传统的 CNN、LSTM、Bi-LSTM 等基准模型,分别提升了 5.6% 、5.1% 和 4.5% 。从表 1 可知,单独以 RGCN 编

码器或者 GAT 编码器完成文本分类的性能低于 BERT 分类模型,但 RGCN 编码器与 GAT 编码器特征融合后的性能却略优于 BERT 分类模型,性能提升了 1.1%。也说明了 RGCN 编码器所提取特征与 GAT 编码器所提取特征具有一定的互补性,使特征融合后能增强模型整体性能。此外,表 3 同时比较了各个模型的大小和推理时间。所提出的异构图神经网络文本分类模型大小与 CNN、LSTM、Bi-LSTM 等模型相近。相比于 BERT 模型(型号为 chinese_L-12_H-768_A-12),所提出模型的模型大小和推理时间远小于 BERT 模型。该模型占用更少的内存空间和计算资源,更有利于模型的在线部署和应用。

表 3　所提出模型与其他深度学习模型的性能对比

模型	宏平均准确率/%	模型大小/MB	推理时间/ms
CNN	53.7	1.32	0.5
LSTM	54.2	1.41	1.31
Bi-LSTM	54.8	3.32	2.36
BERT	58.2	392	139
Model-RGCN+GAT	59.3	1.40	15

4　结束语

该文提出了一种基于异构图神经网络的检务知识咨询业务分类模型。针对用户提问内容长度短的特点,该模型通过 RGCN 编码基于句法依存分析的图表示,并通过 GAT 编码基于邻域窗口的图表示,最后通过特征融合实现了用户提问内容业务类型预测。这种特征融合方式比单独采用 RGCN 编码器的方式提升了 1.9% 的性能,而比单独采用 GAT 编码器的方式提升了 1.5% 的性能。为了平衡这两个编码器的特征学习,分别对 RGCN 编码和 GAT 编码器引入辅助分类器,使模型提升了 1.3% 的性能。为了解决检答网用户提问内容业务类型的样本数据不均衡问题,引入 Focal Loss 损失函数,使模型提升了 5.6% 的性能。此外,所提出模型在测试集上的宏平均准确率均优于传统深度学习文本分类模型,略优于 BERT 分类模型。所提出模型的模型大小和推理时间远小于 BERT 模型。该模型有助于计算机正确理解用户关于检务领域的知识咨询意图,为进一步构建检务智能问答系统提供技术基础。

参考文献:

[1] 赵志刚,金鸿浩. 政法智能化战略下“智慧检务”实践启示[J]. 人民检察,2019(8):29-32.

[2] 高学强. 人工智能时代的中国司法[J]. 浙江大学学报:人文社会科学版,2019,49(4):229-240.

[3] SCHLICHTKRULL M,KIPF T N,BLOEM P,et al. Modeling relational data with graph convolutional networks[C]// European semantic web conference. Switzerland: Springer,2018:593-607.

[4] PETAR V,GUILLEM C,ARANTXA C,et al. Graph attention networks[EB/OL]. (2018-02-04). https://arxiv.org/abs/1710.10903.

[5] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(2):318-327.

[6] CHEN Z,SHI G,WANG X. Text classification based on naive bayes algorithm with feature selection[J]. International Journal on Information,2012(15):4255-4260.

[7] DILRUKSHI I,ZOYSA K D,CALDERA A. Twitter news classification using SVM[C]//8th international conference on computer science & education. Colombo:IEEE,2013:287-291.

[8] 贾俊华. 一种基于 AdaBoost 和 SVM 的短文本分类模型[D]. 天津:河北工业大学,2016.

[9] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Stroudsburg: ACL, 2014:1746-1751.

[10] ZHANG X,ZHAO J,LECUN Y. Character-level convolutional networks for text classification[C]//Advances in neural information processing systems. [s. l.]:ACM,2015:649-657.

[11] ZHANG L,CHEN C. Sentiment classification with convolutional neural networks: an experimental study on a large-scale Chinese conversation corpus[C]//12th international conference on computational intelligence and security. Wuxi: IEEE,2016:165-169.

[12] 郭东亮,刘小明,郑秋生. 基于卷积神经网络的互联网短文本分类方法[J]. 计算机与现代化,2017(4):78-81.

[13] LIU P F,QIU X P,HUANG X J. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the twenty-fifth international joint conference on artificial intelligence. New York:ACM,2016:2873-2879.

[14] ZHOU Y,XU B,XU J,et al. Compositional recurrent neural networks for Chinese short text classification[C]//IEEE/WIC/ACM international conference on web intelligence. [s.

l.]:IEEE,2016:137-144.
[15] ZHOU C,SUN C,LIU Z,et al. A C-LSTM neural network for text classification[EB/OL]. (2015-11-30). https://arxiv. org/abs/1511. 08630.
[16] 和志强,杨 建,罗长玲. 基于 BiLSTM 神经网络的特征融合短文本分类算法[J]. 智能计算机与应用,2019,9(2):21-27.
[17] DEVIN J,CHANG M W,LEE K,et al. Bert:pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24). https://arxiv. org/abs/1810. 04805.
[18] WANG Z N,HUANG Z L,GAO J L. Chinese text classification method based on BERT word embedding[C]//Proceedings of the 5th international conference on mathematics and artificial intelligence. New York:ACM,2020:66-71.
[19] SUN C,QIU X P,XU Y,et al. How to fine-tune BERT for text classification? [C]//China national conference on Chinese computational linguistics. Switzerland: Springer, 2019: 194-206.
[20] 段丹丹,唐加山,温 勇,等. 基于 BERT 模型的中文短文本分类算法[J]. 计算机工程,2021,47(1):79-86.
[21] ZHOU J,CUI G Q,ZHANG Z Y,et al. Graph neural networks:a review of methods and applications [EB/OL]. (2021-04-09). https://arxiv. org/abs/1812. 08434.
[22] XU K L,HU W H,LESKOVEC J,et al. How powerful are graph neural networks? [EB/OL]. (2019-02-22). https://arxiv. org/abs/1810. 00826.
[23] YAO L,MAO C S,LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the AAAI conference on artificial intelligence. Menlo Park:AAAI,2019:7370-7377.
[24] HUANG L Z,MA D H,LI S J,et al. Text level graph neural network for text classification[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing. Stroudsburg:ACL,2019:3444-3450.
[25] JOOST B,IVAN T,WILKER A,et al. Graph convolutional encoders for syntax-aware neural machine translation[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. Stroudsburg:ACL,2017:1957-1967.
[26] LIU X,LUO Z,HUANG H. Jointly multiple events extraction via attention-based graph information aggregation[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. Stroudsburg:ACL,2018:1247-1256.
[27] THIEN H N,RALPH G. Graph convolutional networks with argument-aware pooling for event detection[C]//Proceedings of the 32nd AAAI conference on artificial intelligence. Menlo Park:AAAI,2018:5900-5907.
[28] 王智悦,于 清,王 楠,等. 基于知识图谱的智能问答研究综述[J]. 计算机工程与应用,2020,56(23):1-11.
[29] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-09-07). https://arxiv. org/abs/1301. 3781.
[30] SZEGEDY C,LIU W,JIA Y,et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston:IEEE,2015:1-9.