基于预训练模型的多标签专利分类研究*

佟昕瑀　赵蕊洁　路永和
（中山大学信息管理学院 广州 510006）

摘要：**【目的】**提高专利自动分类效果，准确地为专利申请书匹配适合的一个或多个IPC分类号。**【方法】**构建了大规模中文专利数据集（CNPatents），选取IPC分类号中的前4位作为分类标签，使用BERT、RoBERTa和RBT3模型进行训练和测试。**【结果】**实验结果表明，在含有600多个类别的分类任务中，最好的模型分类准确率为0.756，Micro-F1值为0.597；经过高频标签筛选后，准确率提升到0.912，Micro-F1值提升到0.717。**【局限】**作为训练集的专利文本存在数据不平衡的状况，对训练集进行高频标签筛选仍未完全解决该问题，需要进一步扩大专利数据集规模。**【结论】**实现了多标签专利的自动分类，并通过高频标签筛选进一步提升了模型的分类效果。

关键词：专利分类　预训练模型　专利文本表示
分类号：G350
DOI：10.11925/infotech.2096-3467.2021.0930
引用本文：佟昕瑀，赵蕊洁，路永和. 基于预训练模型的多标签专利分类研究[J]. 数据分析与知识发现，2022，6(2/3)：129-137.(Tong Xinyu, Zhao Ruijie, Lu Yonghe. Multi-label Patent Classification with Pre-training Model[J]. Data Analysis and Knowledge Discovery, 2022, 6(2/3): 129-137.)

1　引　言

专利作为衡量国家创新能力的重要文献，是发现先进技术的重要途径。根据国家知识产权局公布的数据[1]，2020年全国专利申请量达519.4万件，其中发明专利申请149.7万件，实用新型专利申请292.7万件，外观设计专利申请77万件。全年专利授权量为363.9万件，其中高价值专利审查周期平均17.3个月。专利文献的数量快速增长使人工和小规模机器学习的分类方法很难满足现有需求。因此，如何从大量专利申请中快速、准确地识别出专利蕴含的语义信息，并对专利进行准确分类是需要研究的重要内容。

目前国际上常用的专利分类标准主要有国际专利分类（IPC）、欧洲专利分类号（ECLA）、美国专利分类号（UPC/USPC）、日本分类法（FI/F-term）和联合专利分类（CPC）等。比较通用的是国际专利分类（简称IPC），该分类标准根据专利所涉及的不同技术领域，按照IPC标准赋予专利一个或多个分类号，方便对专利文献进行整理和检索，同时根据IPC分类号为专利审查员推送合适的待审专利进行审查。

在2021年1月更新的IPC分类体系中，包含8个部、131个大类、646个小类、7 523个大组和68 899个小组，共计76 422个组别。每个专利可以被赋予多个分类号，且类别之间的相似程度较高，这给专利分类工作带来巨大挑战。专利申请人或专利审查员在对专利进行分类时，需要从大量概念相近的类别中选择恰当的分类号，这要求专利申请人或专利审查员对IPC分类体系有深度了解，并通过阅读大量专

通讯作者（Corresponding author）：路永和（Lu Yonghe），ORCID：0000-0002-7758-9365，E-mail：luyonghe@mail.sysu.edu.cn。

*本文系广东省重点领域研发计划项目（项目编号：2021B0101420004）和广东省区域联合基金重点项目（项目编号：2019B1515120085）的研究成果之一。

The work is supported by the Research and Development Program in Key Areas of Guangdong Province of China (Grant No. 2021B0101420004), the Regional Joint Fund Key Projects of Guangdong Province of China (Grant No. 2019B1515120085).

利文献进行判断。虽然现在已有深度学习方法应用在专利自动分类中,但多数集中在主分类号的单一分类研究上,且选择的分类号粒度也停留在部或是大类级别,对细粒度的专利分类无法提供有效帮助。

因此,在已有研究的基础上,本文通过构建大规模专利分类数据集,使用预训练好的BERT、RoBERTa和RBT3深度学习模型,针对多标签专利分类任务进行微调,使用Sigmoid作为激活函数,使用BCEWithLogitsLoss作为损失函数,构建多标签专利分类模型,采用准确率(Accuracy)和Micro-F1值作为评价指标,验证预训练模型在大规模数据集上的多标签专利分类效果。本文的创新之处体现在以下方面。

(1)构建可以用于多标签专利分类任务的数据集,且数据集规模可扩展。

(2)微调BERT、RoBERTa和RBT3模型,使其可以适应多标签专利分类任务。

(3)将专利分类粒度精确到"小类"级别,且构建的数据集包含了所有小类标签。

(4)通过高频标签筛选,本文所构造的多标签专利分类模型的准确率和Micro-F1值分别达到91.2%和71.7%。

2 相关研究

2.1 专利文本表示模型

文本表示一直是文本信息自动处理中不可缺少的关键技术之一。2016年中国中文信息学会(Chinese Information Processing Society of China,CIPS)发布的《中文信息处理发展报告(2016)》将语言表示方法按表示形式划分为离散表示和连续表示,按不同粒度划分为字、词、句子和篇章等不同层次的表示[2]。本文主要通过深度学习算法将专利文本映射到一个低维连续空间,进而完成专利分类任务。因此,本节主要介绍近年来文本表示技术的基本理论和发展情况。

词嵌入最常采用的实现手段是2013年提出的Word2Vec技术[3],包括连续词袋模型(Continuous Bag-of-Words Model,CBOW)和Skip-gram模型。在Word2Vec发布不久后,斯坦福自然语言处理组在2014年提出Global Vectors for Word Representation(GloVe)[4]。之后,Peters等[5]提出的Embeddings from Language Models(ELMo)方法通过构建深层双向长短期记忆(Long Short-Term Memory,LSTM)模型获得词语的向量表示,借助多任务思想,从低层到高层逐步解决从语法到语义、语境方面的特征捕捉,有效解决了一词多义问题。由于LSTM的特征抽取能力远弱于Transformer[6],GPT(Generative Pre-Training)则使用Transformer特征抽取器,利用基础语言模型进行预训练,然后通过Fine-tuning模式解决下游任务。基于ELMo和GPT模型,谷歌公司于2018年提出的BERT预训练模型[7],其泛化能力极强,当时在11个任务中Fine-tuning均取得state-of-the-art的性能。Liu等[8]重新研究了BERT的预训练机制,评估了超参数和训练集大小对BERT模型的效果影响。发现BERT的预训练并不充分,据此提出了改进BERT的方法,包括增大Batch Size并在更大的数据集上对BERT进行训练、不再使用NSP(Next Sentence Prediction)任务、使用更长的序列进行训练、动态调整Masking机制以及使用更大的Byte-Level BPE。经过实验发现,重新训练后的RoBERTa模型在相同的任务上性能相较BERT模型有了进一步的提升。

综上所述,训练任务的选择会对获得的句子嵌入的质量造成影响。单一任务训练的句子嵌入只能揭示句子中一部分语义,而不同任务结合训练得到的句子嵌入能揭示句子不同层面的语义信息。目前,预训练模型已经成为主流的文本表示模型,在许多任务上都超越了传统的词嵌入方法。

2.2 专利分类研究

专利的分类检索系统是专利系统重要组成部分,一方面通过分类检索系统能够有效地节约研究者的搜集时间,另一方面也能通过精准的检索为研究者提供最新的研究方向。传统的专利分类模型是采用机器学习方法,分类速度较快,但是准确率不能满足需求。而深度学习能够应对复杂文本类型的分类,提高分类准确率和分类效率。经过检索发现,近年专利分类采用的分类器以深度学习模型为主,选用的数据主要是英文专利数据。

针对专利类别由粗到细可划分为不同级别这一特征,Kowsari等[9]提出基于深度学习的层次文本分

类架构，在不同级别的分类中使用 RNN、CNN 和 DNN 这三种不同的分类模型。实验证明，基于深度学习的层次分类方法在三个专利数据集的分类效果普遍优于传统方法。李生珍等[10]使用向量空间模型对专利的标题和摘要进行表示，经过卡方统计量筛选特征后输入 BP 神经网络划分专利所属的 IPC 类别。Xiao 等[11]用 VSM 对专利文本进行稀疏表示，通过稀疏自编码器对稀疏特征向量进行压缩，然后采用深度置信网络进一步提取文本深层特征，并输入 Softmax 分类器获得专利文本的类别。马双刚[12]在对文本进行表示和特征筛选后，输入降噪自编码器（Denoising Auto Encoder，DAE），进一步抽取特征获取专利文本的低维表示，最后用 SVM 对专利分类。胡杰等[13]将词向量输入 CNN 进行专利文本表示，再用随机森林算法对专利类别进行预测。

上述研究都是针对单标签专利文本的分类研究。而在多标签专利文本的分类研究方面，包翔等[14]结合专利文本的固有格式以及每个专利文本可以拥有多个 IPC 分类号的情况，将多示例多标签学习应用于专利自动分类中。吕璐成等[15]考虑到传统机器学习方法存在的缺陷，综合考虑专利文本语序特征、上下文特征和分类关键特征，设计 Word2Vec+TextCNN、Word2Vec+GRU 等 7 种深度学习模型，选取 IPC 主分类号的“部”作为分类依据，实验证明深度学习模型实验效果优于传统分类模型。Li 等[16]提出一种能够应用于大型专利分类的深度学习算法 DeepPatent，选取美国国家专利数据作为数据集，用标题和摘要代表专利文本主要信息，然后使用 Skip-gram 模型将文本单词转换为词向量，再将词向量连接成密集矩阵，最后将矩阵输入 CNN 模型进行多标签专利文本分类。Lee 等[17]在 DeepPatent 的基础上提出了 USPTO-3M 数据集，使用权利要求书作为专利文本的主要内容，用预训练的 BERT 模型作为主要模型，实验证明 BERT 模型针对多标签的专利文本有更好的分类效果。

综上所述，英文专利分类已经有比较成熟的数据集和研究内容，中文专利目前缺少大规模数据集作为研究支撑，且分类号精确粒度较粗。因此，本文通过构建大规模的中文专利数据集用以进行多标签专利分类任务，选取 IPC 分类号的前 4 位作为分类标签，采用预训练模型进行训练和测试。

3　相关模型和技术

3.1　BERT 模型

BERT 模型是由谷歌团队于 2018 年提出的预训练模型[7]，全称是 Bidirectional Encoder Representation from Transformers，即双向 Transformer 的文本表示模型，使用了 Masked LM 和 Next Sentence Prediction 两种方法分别捕捉词语和句子级别的语义信息。

3.2　Transformer 模型

2017 年，谷歌团队提出了 Transformer 模型解决 Sequence-to-Sequence 问题[6]。该模型放弃了传统 Encoder-Decoder 框架与 CNN 或 RNN 模型相结合的固有方式，使用 Attention 结构并行训练模型，大大减少了模型的计算量，提高了并行计算的效率，在多个数据集上取得了良好的实验结果。图 1 为 Transformer 模型的结构示意图，它由左边的编码模块（Encoder）和右边的解码模块（Decoder）组成。

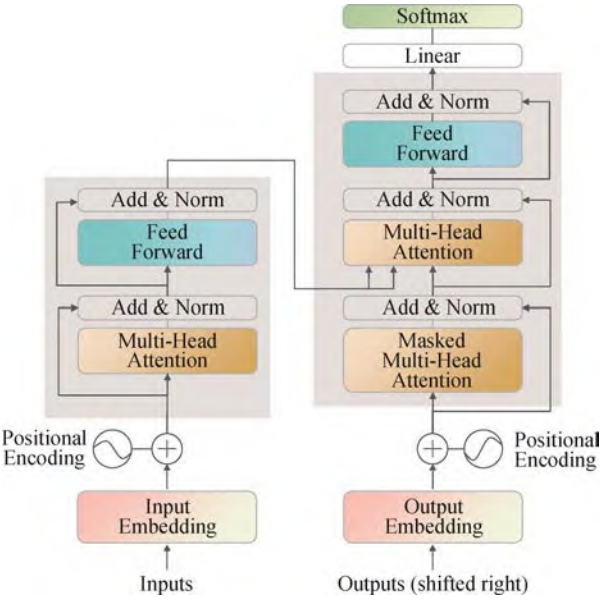

图 1　Transformer 模型示意图[6]

Fig.1　Diagram of the Transformer Model

在编码器模块中，编码器的编码主要由多头自注意力机制（Multi-Head Attention）、全连接前馈神经网络层（Feed Forward）及残差连接和归一化（Add & Norm）构成。首先，在输入文本被嵌入后，通过结合编码位置信息得到文本的最终输入向量矩阵；其

次，在自注意力机制中，每个单词有三个不同的向量表示，分别是Query向量（$\boldsymbol{Q}$）、Key向量（$\boldsymbol{K}$）和Value向量（$\boldsymbol{V}$），长度均为64，可以根据文本嵌入向量得到$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$三个向量；最后，对文本向量矩阵进行不同的线性变化，得到$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的不同组合，再将$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的不同组合的自我注意机制的输出相加，得到多头注意机制的潜在输出。通过这种方式，模型可以学习不同方面的文本表示特性，其计算结果如公式（1）所示。在加入残差连接和归一化之后，输入被送入全连接层进行非线性转换，最后得到潜在语义文本表示的输出。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (1)$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别代表Query向量、Key向量和Value向量，d_k代表向量维度。

在解码器模块中，其输入是编码器层的输出和解码器层在相应的前一个位置的输出。因此，它的$\boldsymbol{K}$和$\boldsymbol{V}$来自编码器层的输出，而它的$\boldsymbol{Q}$来自解码器层在相应的前一个位置的输出。解码器模块的输出是相应位置上输出字的概率分布。

3.3 词嵌入

BERT模型处理输入文本中的单词向量嵌入方式如图2所示。可以看出，BERT的词嵌入表示包括词嵌入（Token Embedding）、句子嵌入（Segment Embedding）和位置嵌入（Position Embedding）。图2所示的例子是BERT对任务的句子的编码，除了文本内容外，还包含[CLS]和[SEP]两个特殊标记。[CLS]置于句子对的开头，是用于分类任务的标志，该标志的最终结果通常被用作表示整个句子或句子对的向量；[SEP]是用于分离输入文本中两个句子的分离标志，在句子对中起重要作用。

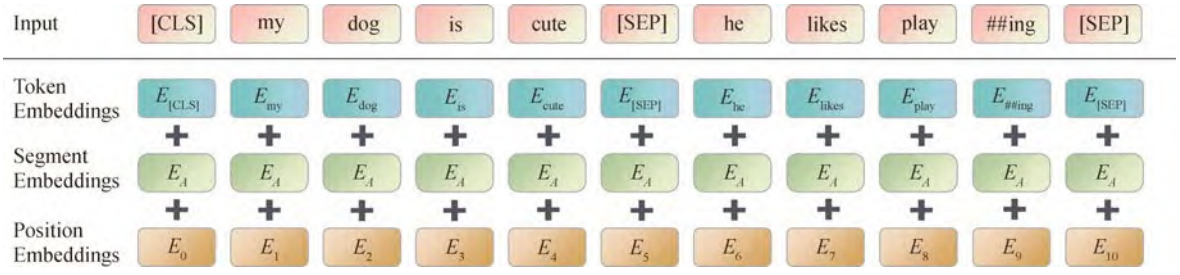

图2 BERT对输入文本的词向量嵌入处理[7]

Fig.2 Word Vector Embedding Processing of Input Text by BERT

3.4 全词遮盖

为了训练文本深度双向表示，Devlin等[7]选择直接随机地遮盖住一定比例的输入标记，然后仅预测被遮住的输入标记，这种方式被称为Masked LM（MLM）。在这种情况下，和其他标准语言模型相同，被遮盖的标记对应的最终隐藏向量被当作Softmax的关于该词的一个输出。

但是由于中文与英文句子的组成方式不同，谷歌官方发布的BERT-base-Chinese中，中文以字为粒度进行切分，没有考虑到传统NLP中文分词问题，在使用同样的方法对中文文本进行遮盖时，会出现一个词组只有一个字被遮盖的情况，使词组失去了原本的意思。因此，在预训练阶段，使用全词遮盖（Whole Word Masking）作为预训练阶段的训练样本生成策略[18]。在全词遮盖中，如果一个完整的词的部分子词被遮盖，则同属该词的其他部分也会被遮盖。全词遮盖的生成样例如表1所示。

表1 全词遮盖生成样例[18]

Table 1 Full Word Masking Generation Example

说明	样例
原始文本	使用语言模型来预测下一个词的probability。
分词文本	使用 语言 模型 来 预测 下 一个 词 的 probability 。
原始遮盖输入	使 用 语 言 [MASK] 型 来 [MASK] 测 下 一 个 词 的 pro [MASK] ##lity 。
全词遮盖输入	使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK] [MASK] 。

4 基于预训练模型的多标签专利分类方法

4.1 实验架构

本文对BERT、RoBERTa和RBT3模型进行微

调，使模型能够适应多标签专利分类任务。实验架构如图3所示。

图3　基于预训练模型的多标签专利分类实验架构

Fig.3　Architecture for Multi-Label Patent Classification Based on Pre-Trained Models

4.2　数据采集与预处理

本文实验数据来源为Google Patent网站2018-2021年的中文专利文本，为保证多标签实验的广泛性，故不指定特定领域，共随机采集到1 033 917份专利文档，每篇文档保留标题、摘要和分类号信息，选择分类号前4位作为类别标签，并将采集到的专利文本分为两个数据集：CNPatents-Large（包含采集的全部专利文本）和CNPatents-Small（包含采集的部分专利文本），按照8∶2的比例设置训练集和测试集。数据集具体信息如表2所示。

此外，由于两个数据集都存在部分标签下的训练样本少于100篇的情况，会影响预训练模型对该类别专利文本进行特征学习的效果。因此，对每个类别标签下的专利文本数量进行统计，并筛选专利文本数量排名前30的标签作为高频标签。筛选后的类别标签及数据集详情如表3所示，两个数据集排名前30的标签及其对应的文本数如表4和表5所示。

表2　数据集详情

Table 2　Dataset Details

数据集名称	专利文本数量	标签数量	训练集	测试集
CNPatents-Large	1 033 917	654	827 134	206 783
CNPatents-Small	398 527	638	318 822	79 705

表3　高频标签筛选后的数据集详情

Table 3　Details of the Dataset after High-Frequency Tag Filtering

数据集	专利文本量	标签数量	训练集	测试集
CNPatents-Large（30）	685 133	30	548 106	137 027
CNPatents-Small（30）	314 424	30	251 539	62 885

表4　高频标签筛选后的CNPatents-Large（30）数据集详情

Table 4　Details of the CNPatents-Large（30） Dataset after High-Frequency Label Filtering

分类标签	文本数/篇	分类标签	文本数/篇	分类标签	文本数/篇
G06F	43 183	H04W	12 522	G01R	7 086
Y02E	28 405	Y02P	12 126	H02J	7 086
G06K	26 263	H01M	9 922	G01S	6 809
H04L	24 516	Y02A	9 625	G05B	6 506
G06Q	23 626	Y02T	8 473	C08L	6 494
G01N	17 212	B01D	8 201	Y02B	6 484
G06N	15 570	A61K	8 191	C04B	6 410
G06T	15 059	C02F	7 370	B01J	6 340
H01L	13 611	A61B	7 215	C22C	6 288
H04N	12 730	C08K	7 120	G02B	6 189

表5　高频标签筛选后的CNPatents-Small（30）数据集详情

Table 5　Details of the CNPatents-Small（30） Dataset after High-Frequency Label Filtering

分类标签	文本数/篇	分类标签	文本数/篇	分类标签	文本数/篇
G06F	12 344	H04W	2 819	G01R	1 740
G06K	9 209	Y02P	2 554	A61K	1 714
G06Q	7 293	H01M	2 537	G08G	1 681
Y02E	7 054	Y02T	2 226	Y02B	1 654
G06N	6 566	Y02A	2 196	H02J	1 653
H04L	6 534	B01D	1 931	G05D	1 637
G06T	4 885	G01S	1 883	G05B	1 635
G01N	4 110	A61B	1 825	B25J	1 618
H01L	3 725	C22C	1 793	F24F	1 599
H04N	3 616	B08B	1 792	G02B	1 560

4.3　预训练模型

本文选择由哈工大讯飞联合实验室训练的

BERT-wwm-ext、RoBERTa-wwm-ext和RBT3作为实验模型[18]。其中，RoBERTa-wwm-ext是按照RoBERTa训练方式训练出的BERT模型，RBT3是由RoBERTa-wwm-ext 3层进行初始化，继续训练了100万步所得到的模型，模型具体参数如表6所示。

表6 本文使用的模型具体数据

Table 6 Specific Data of the Model Used in This Paper

模型 / 属性	BERT-wwm-ext	RoBERTa-wwm-ext	RBT3
词汇遮盖方式	Whole Word Masking	Whole Word Masking	Whole Word Masking
原始模型	BERT-base	BERT-base	BERT-base
数据来源	中文维基百科，其他百科、新闻、问答等数据，总词数达54亿。	中文维基百科，其他百科、新闻、问答等数据，总词数达54亿。	中文维基百科，其他百科、新闻、问答等数据，总词数达54亿。
训练步长	$1M^{MAX128}+400K^{MAX512}$	$1M^{MAX512}$	$1M^{MAX512}+1M^{MAX512}$
训练集样本数量	2,560 / 384	384	384
优化器	LAMB	AdamW	AdamW
词汇表	21 128	21 128	21 128

4.4 分类任务

多分类任务与单分类任务不同，在计算各类别概率时需要对不同类别的概率值进行独立计算，无法使用Softmax函数作为分类任务的激活函数。因此，本实验在输出层使用的激活函数及对应的损失函数如表7所示。

表7 分类任务使用的激活函数及损失函数

Table 7 Activation Functions and Loss Functions Used for the Classification Task

分类任务名称	输出层使用的激活函数	对应的损失函数
多标签分类	Sigmoid()	BCEWithLogitsLoss()

Sigmoid函数多用于二分类或多标签分类任务，该函数输出每维代表样本是否属于该类别的概率，对所有维求和的结果不等于1，因此可以对n个类别分别计算概率，适用于多标签分类任务，如公式(2)所示。

$$\mathrm{Sigmoid}(z_i)=\frac{1}{1+e^{z_i}} \tag{2}$$

使用Sigmoid函数得到输出后，BCELoss函数在每个类维度上求交叉熵损失然后加和求平均，得到最后的分类结果。

BCEWithLogitsLoss将Sigmoid函数和BCELoss函数整合起来，比纯粹使用BCELoss+Sigmoid更加稳定，如公式(3)所示。

$$\mathrm{loss}(z,y)=\frac{1}{N}\sum_{n=1}^{N}-w_n(y_n\log(\delta(z_n))+(1-y_n)\log(1-\delta(z_n))) \tag{3}$$

其中，z_n表示预测第n个样本为正例的得分(尚未经过Sigmoid处理)；y_n表示第n个样本的标签；$\delta()$表示Sigmoid函数。

5 实验与分析

5.1 实验环境

本文实验环境如表8所示。

表8 实验环境配置参数

Table 8 Experimental Environment Configuration Parameters

环境	配置参数
处理器	INTEL XEON GOLD 6139M（2.3~3.7 GB）
显卡	NVIDIA GeForce RTX 2080 Ti
内存	8 × 11 GB
操作系统	Ubuntu 16.04 64bit
语言	Python

5.2 实验参数及评价指标

(1) 实验参数

用于训练和测试的数据集文本组合为专利标题+摘要，在表8所示的实验环境中进行多次实验，分别对下列参数进行微调，选取实验效果最好的一组参数作为正式实验使用的参数，具体参数如表9所示。

表9　实验参数设置

Table 9　Experimental Parameter Settings

参数	设定值
MAX_LEN	200
TRAIN_BATCH_SIZE	16
VALID_BATCH_SIZE	16
EPOCHS	3
LEARNING_RATE	1e-5

（2）评价指标

常用分类模型评价指标是根据混淆矩阵进行计算而来的，如表10所示。

表10　模型评价的混淆矩阵

Table 10　Confusion Matrix for Model Evaluation

真实值 / 预测值	Positive	Negative
Positive	TP（预测结果为正的正样本）	FP（预测结果为正的负样本）
Negative	FN（预测结果为负的正样本）	TN（预测结果为负的负样本）

本文选取准确率（Accuracy）和Micro-F1值作为模型评价指标，准确率指分类预测正确的样本数占总样本数的比例，如公式（4）所示。

$$Accuracy = \frac{TP + TN}{P + N} \tag{4}$$

Micro-F1值考虑$Precision_{micro}$和$Recall_{micro}$的综合性指标，如公式（5）-公式（7）所示。

$$Precision_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{5}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{6}$$

$$Micro\text{-}F1 = 2 \cdot \frac{Precision_{micro} \cdot Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{7}$$

Micro-F1值是衡量多标签分类任务效果的重要指标，对同一类别下的文本给予相同权重，分别计算每类别的精确率（Precision）和召回率（Recall）并求F1值，最后求所有类别的平均F1值，这种处理方式的优势在于对采样数量少的类别关注更多，可以部分消除因为数据不平衡带来的影响。

5.3　实验结果与分析

（1）基于预训练模型的专利分类实验

使用微调后的BERT-wwm-ext、RoBERTa-wwm-ext和RBT3模型在CNPatents-Large和CNPatents-Small数据集上分别进行实验，结果如表11所示。

表11　多标签专利分类实验结果

Table 11　Results of Multi-Label Patent Classification

模型 \ 数据集及指标	CNPatents-Large		CNPatents-Small	
	准确率	Micro-F1	准确率	Micro-F1
BERT-wwm-ext	0.659	0.597	0.756	0.506
RoBERTa-wwm-ext	0.657	0.594	0.746	0.470
RBT3	0.646	0.567	0.736	0.439

在两个数据集上分别进行实验，效果最好的是BERT-wwm-ext模型，其中在CNPatents-Small数据集上的准确率达到0.756，高于CNPatents-Large的0.659，但是在CNPatents-Large上的Micro-F1值要优于另一数据集，达到0.597，这是因为预训练模型依赖训练集的数据量。有实验证明，增大预训练模型的训练语料，可以提高预训练模型在任务中的准确率等指标[18]。

（2）高频标签筛选后的专利分类实验

对数据集进行高频标签筛选后重复前述实验，实验结果如表12所示。

表12　高频标签筛选后的实验结果

Table 12　Results after High-Frequency Label Screening

模型 \ 数据集及指标	CNPatents-Large		CNPatents-Small	
	准确率	Micro-F1	准确率	Micro-F1
BERT-wwm-ext	0.862	0.717	0.912	0.693
RoBERTa-wwm-ext	0.863	0.717	0.912	0.696
RBT3	0.860	0.707	0.910	0.669

经过高频标签筛选后的模型分类效果得到显著提升，在CNPatents-Large数据集上，准确率达到0.863，Micro-F1值提升到0.717；在CNPatents-Small数据集上，准确率为所有模型和数据集中的最好效果，达到0.912，Micro-F1值提升到0.696。可见，对数据集进行平衡处理可以有效地提高预训练模型的分类效果。

6　结　语

为了改善专利自动分类效果，为专利申请书匹配适合的一个或多个IPC分类号。首先，本文获取了2018-2021年的中文专利文本100万余份，构建了

用于多标签专利分类任务的数据集CNPatents-Large和CNPatents-Small;其次,提出一种基于预训练的多标签专利分类方法,在构建的专利数据集上进行训练,对BERT、RoBERTa和RBT模型进行微调,实现大规模多标签专利的自动分类;最后,通过高频标签筛选进一步提升了模型的分类效果,其中准确率提升了0.107,Micro-F1值提升了0.120。本文后续研究可以从以下两方面开展:

(1)扩大专利分类数据集规模,增加每个分类号下的训练样本数量,使预训练模型可以充分学习不同分类号下专利文本的特征,进一步提高专利自动分类效果;

(2)细化用于训练的专利分类号的粒度,研究精确到大组或小组级别的专利自动分类模型。

参考文献:

[1] 2020知识产权统计年报. 分国内外三种专利申请/授权/有效量(2020年)[R]. 国家知识产权局, 2020. (2020 Annual Report on Intellectual Property Statistics. Three Types of Patent Applications/Grants/Validities by Domestic and Foreign Countries (2020) [R]. China National Intellectual Property Administration, 2020.)

[2] 中国中文信息学会. 中文信息处理发展报告(2016)[R]. 北京, 2016. (Chinese Information Processing Society of China. Report on the Development of Chinese Information Processing(2016) [R]. Beijing, 2016.)

[3] Mikolov T, Chen K, Corrado G S, et al. Efficient Estimation of Word Representations in Vector Space [OL]. arXiv Preprint, arXiv: 1301.3781.

[4] Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[5] Peters M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 2227-2237.

[6] Vaswania A, Shazeer N, Parmar N, et al. Attention is All You Need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017: 6000-6010.

[7] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [OL]. arXiv Preprint, arXiv:1908.08962v2.

[8] Liu Y H, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [OL]. arXiv Preprint, arXiv: 1907.11692.

[9] Kowsari K, Brown D E, Heidarysafa M, et al. HDLTex: Hierarchical Deep Learning for Text Classification [C]// Proceedings of 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017: 364-371.

[10] 李生珍, 王建新, 齐建东, 等. 基于BP神经网络的专利自动分类方法[J]. 计算机工程与设计, 2010, 31(23): 5075-5078.(Li Shengzhen, Wang Jianxin, Qi Jiandong, et al. Automated Categorization of Patent Based on Back-Propagation Network [J]. Computer Engineering and Design, 2010, 31(23): 5075-5078.)

[11] Xiao L Z, Wang G Z, Zuo Y. Research on Patent Text Classification Based on Word2Vec and LSTM[C]//Proceedings of the 11th International Symposium on Computational Intelligence and Design (ISCID). 2018: 71-74.

[12] 马双刚. 基于深度学习理论与方法的中文专利文本自动分类研究[D]. 镇江: 江苏大学, 2016.(Ma Shuanggang. The Study of Automatic Chinese Patent Classification Based on Deep Learning Theory and Method[D]. Zhenjiang: Jiangsu University, 2016.)

[13] 胡杰, 李少波, 于丽娅, 等. 基于卷积神经网络与随机森林算法的专利文本分类模型[J]. 科学技术与工程, 2018, 18(6): 268-272.(Hu Jie, Li Shaobo, Yu Liya, et al. A Patent Classification Model Based on Convolutional Neural Networks and Rand Forest [J]. Science Technology and Engineering, 2018, 18(6): 268-272.)

[14] 包翔, 刘桂锋, 崔靖华. 多示例多标签学习在中文专利自动分类中的应用研究 [J]. 图书情报工作, 2021, 65(8): 107-113.(Bao Xiang, Liu Guifeng, Cui Jinghua. Application of Multi Instance Multi Label Learning in Chinese Patent Automatic Classification [J]. Library and Information Service, 2021, 65(8): 107-113.)

[15] 吕璐成, 韩涛, 周健, 等. 基于深度学习的中文专利自动分类方法研究 [J]. 图书情报工作, 2020, 64(10): 75-85.(Lyu Lucheng, Han Tao, Zhou Jian, et al. Research on the Method of Chinese Patent Automatic Classification Based on Deep Learning [J]. Library and Information Service, 2020, 64(10): 75-85.)

[16] Li S B, Hu J, Cui Y X, et al. DeepPatent: Patent Classification with Convolutional Neural Networks and Word Embedding[J]. Scientometrics, 2018, 117(2): 721-744.

[17] Lee J S, Hsiang J. Patent Classification by Fine-Tuning BERT Language Model [J]. World Patent Information, 2020, 61: 101965.

[18] Cui Y M, Che W X, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

作者贡献声明：

佟昕瑀：采集、清洗、处理数据，进行实验，论文起草；
赵蕊洁：论文最终版本修订；
路永和：提出研究思路，设计研究方案。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail: tongxy7@mail2.sysu.edu.cn。
[1] 佟昕瑀. Dataset.zip. 专利数据集.
[2] 佟昕瑀. Code.zip. 论文实验源代码.
[3] 佟昕瑀. 实验结果.docx. 实验结果数据.

收稿日期：2021-08-30
收修改稿日期：2021-11-15

Multi-label Patent Classification with Pre-training Model

Tong Xinyu　Zhao Ruijie　Lu Yonghe
(School of Information Management, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract: [**Objective**] This paper tries to improve the automatic patent classification method and accurately match patent applications with one or more suitable IPC classification numbers. [**Methods**] We constructed a large-scale Chinese patent dataset (CNPatents), and used the first four digits of IPC classification numbers as labels. Then, we utilized BERT, RoBERTa, and RBT3 models for training and testing. [**Results**] For our classification task with more than 600 labels, the best model reached an accuracy of 75.6% and a Micro-F1 value of 59.7%. After high-frequency label screening, the accuracy and the Micro-F1 value increased to 91.2% and 71.7%. [**Limitations**] The patent documents as the training set have extreme data imbalance issue, which needs more research to improve the high-frequency tag screening for the training. [**Conclusions**] This paper realizes the automatic classification of multi-label patents and further improves the performance of classification model with high-frequency label screening.

Keywords: Patent Classification　Pre-Training Model　Patent Text Representation