计算机科学
COMPUTER SCIENCE

接诉即办智能派单业务调度算法研究

贾经冬, 张敏南, 赵祥, 黄坚

引用本文

贾经冬, 张敏南, 赵祥, 黄坚. 接诉即办智能派单业务调度算法研究[J]. 计算机科学, 2023, 50(11A): 230300029-7.

JIA Jingdong, ZHANG Minnan, ZHAO Xiang, HUANG Jian. Study on Scheduling Algorithm of Intelligent Order Dispatching [J]. Computer Science, 2023, 50(11A): 230300029-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300029

接诉即办智能派单业务调度算法研究

贾经冬　张敏南　赵　祥　黄　坚
北京航空航天大学软件学院　北京 100191

摘　要　随着国家数字化建设的发展,社会治理的智能化、专业化也成为城市科技进步的基本要求,各政府系统须要对人民的诉求做到高效精确的处理。而从当前的各大政府门户网站的诉求通道收集的民众诉求信息,均是通过人工方式判断责任部门,然后将其手动分配给相关部门进行后续问题的核实和处理,大大限制了诉求处理的效率和准确性。而接诉即办智能派单算法利用人工智能和深度学习方法,基于真实的民众诉求信息数据进行训练,自动精准而高效地将诉求分派到相关部门进行后续审查处理,加快了政务处理流程的速度并大大降低了不必要的人力成本,因此该智能调度算法的研究有着重要意义。首先,通过数据去噪和脱敏,将数据进行层级拼接,构建数据标签和标准流程库以进行标签对齐。然后,基于公开数据集训练地址识别基线模型,在工单分类中提出基于类别比例采样的标签融合方法解决数据类不平衡问题,实验结果显示在基线模型的基础上提高了数十个百分点。最后,结合分类模型和地址识别模型,构建智能回复模板,完成接诉即办智能派单的全流程。

关键词:智能派单;类不平衡;标签融合;BERT 模型;深度学习
中图法分类号　TP311

Study on Scheduling Algorithm of Intelligent Order Dispatching

JIA Jingdong,ZHANG Minnan,ZHAO Xiang and HUANG Jian
School of Software,Beihang University,Beijing 100191,China

Abstract　With the development of national digital construction,the intellectualization and specialization of social governance have become the basic requirements for the progress of urban science and technology. All government systems need to deal with the demands of people efficiently and accurately. However,the public appeal information collected from the appeal channels of major government portals is manually judged by responsible departments and then manually assigned to relevant departments for follow-up verification and processing,which greatly limits the efficiency and accuracy of appeal processing. Using artificial intelligence and deep learning methods,the intelligent dispatching algorithm based on real public demand information data training,accurately and efficiently dispatches demands to relevant departments,accelerates the speed of government affairs processing process and greatly reduces unnecessary labor costs. Therefore,the research of this scheduling algorithm is of great significance. First,the data is denoised and desensitized,and hierarchical stitching is used to build data labels and standard process libraries for label alignment. Then,a baseline model for address recognition is trained on publicly available datasets,and a label fusion method based on category proportion sampling is proposed to solve the problem of imbalanced data in work order classification. Experimental results show that the method improves the baseline model by varying degrees. Finally,combining the classification model and the address recognition model,an intelligent response template is constructed to complete the entire process of intelligent dispatching for complaint handling.

Keywords　Intelligent order dispatch,Class imbalance,Label combination,BERT model,Deep learning

1　引言

随着数字化时代的来临,与日常生活息息相关的各项服务都开始思考减少不必要的人力成本,转而投向更智能、更高效的服务提供方式。而对于互联网时代呈爆炸式增长的诉求信息,都需要将其与服务责任方一一对应,以使客户的问题得到妥善解决,智能派单技术也就应运而生。

智能派单,即利用数字化时代的数据、人工智能等方法设计出高效而又精确的算法,以更精准地分派工单到负责方,更迅速地响应诉求。目前,其已被作为研究重点应用于多个实践领域。

外卖平台逐步从骑手抢单模式过渡到智能派单模式,将派单场景和突发因素考虑在内,从而在提升骑手效率的同时也使骑手的收益尽可能最大化[1]。对于通信缺陷的告警工单分派[2],以派单响应时间为优化重点,利用时间约束和路径优化方法,设计了通信缺陷智能派单方法。在银行系统中,Liu[3]提出了基于改进 TF-IDF 加权的 word2vec 词嵌入表示和卷积神经网络结合的银行智能派单系统,各项评估指标均

基金项目:国家重点研发计划项目:政法智能协同技术支撑体系与应用示范研究(2020YFC0833400)
This work was supported by the National Key R&D Program of China:Research on the Technical Support System and Application Demonstration of Intelligent Cooperation in Politics and Law(2020YFC0833400).
通信作者:贾经冬(jiajingdong@buaa.edu.cn)

得到了提升。由此可看出，智能派单已在生活中的各项服务得到应用，然而其还没有应用于与民生深度关联的政务系统。

城市治理是实现国家治理体系现代化的重要部分。数字化建设为以往的群众诉求收集提供了新接口，使得人民群众可以通过政府网站的诉求通道直达相关政府，使群众诉求得到妥善处理。然而，随之而来的则是诉求信息的爆炸式增长和日益增高的对逐个诉求的人力判断成本，因此对诉求工单的智能分派被提上日程，以期更高效、更精准地处理民众诉求。

传统的工单派发主要依赖于业务人员对业务的熟练程度，越熟练的业务人员能在越短的时间内准确判断工单的所属类别，并交由相关部门进行后续处理。而在数据化和智能化水平越来越高的今天，仅仅依靠人力带来的问题是效率低下且成本高，因此迫切需要一个更好的解决方案。接诉即办智能派单就是通过构建自动化流程，抽取接收到的市民投诉工单的关键信息，主要包括对工单内容进行分类以及抽取工单中的地址信息，通过地址匹配和回复模板的构建，将市民的投诉工单准确派发到相关部门，由该部门受理此工单，减少繁琐的部门间流转，从而解决目前政务服务的低效运作问题。

本文兼顾企业建设城市大脑的需求，对目前北京市海淀区“一网统管”业务中的接诉即办智能派单业务进行了深入分析研究，提出了解决方案。主要贡献包括以下 3 个方面：

1）本文提出了一种改进于 Bert 模型的智能派单调度算法，以真实政务数据为基础，加快了政务处理的速度；

2）为了解决数据中存在的类不平衡问题，本文提出了基于比例采样的标签融合方法，此方法有效提高了模型的泛化能力；

3）除了对工单进行地址识别和分类以外，本文还对工单的回复模板进行了构建，对输入工单文本给出了回复建议，进一步加强了诉求信息的处理效率。

2 相关工作

在大数据时代，数据的指数级增长使得传统的人工处理方式逐渐落后，因此需要结合人工智能技术使处理过程智能化、便捷化、准确化，许多领域开始引入智能化技术，辅助工单派发。

随着外卖和快递行业的发展壮大，顾客的需求量和骑手的配送负担也同步提升，因此如何更智能地进行派单成了重要难题。Xiong[1]以美团和饿了么为研究对象，深入剖析了外卖平台智能派单的动态优化机理，为外卖平台改进其智能派单系统提供了重要参考。外卖和出行平台的派单智能化主要从配送路径和订单分配两方面进行优化。既有研究中有以客户时间满意度[4]、平台总成本[5]、同时考虑客户优先级和满意度[6]为目标函数的，也有将时间惩罚成本作为变动成本的修正目标函数[7]。在订单分配方面，Ozkan[8]将顾客和司机耐心作为因素加入考量，构建了总服务乘客数最大化的优化函数；Braverman[9]以系统中的空载车辆为核心指标构建优化函数；Xu[10]直接以收益为指标，以当前和未来收益最大化构建优化函数。除此之外，还发展出基于强化学习[10]和基于动态图的算法[11]进行订单的智能化分配。

而在城市治理方面，对于智能化派单的应用少之又少，目前的智能政务系统大多仅仅是利用了互联网和大数据进行收集，而对智能处理的体现不足。上海市建立了三级平台、五级应用的基本架构，力争在最短时间、以相对最小成本解决最突出问题，实现最佳综合效应[12]。南通市以“市域治理”为抓手，建成了推进数据共享和市县镇三级联动的市域治理体系[13]。此外还有浙江省“最多跑一次”的机制，利用政务大数据平台提升治理效率[14]。北京市通过网格化城市管理平台，提供了统一开放的监督检查服务，为市、区、乡镇、社区等相关部门和公众提供了数据共享、监督监管、综合考评、宣传发布等服务[15]，提高了城市公共资源配置的优化能力，初步构建成以“吹哨报到”和“接诉即办”两种体制为核心的城市大脑平台。可以看到，目前在城市治理方面已经出台相关政策以及更新架构，然而对于后续的处理流程大多仍旧依靠人工，并没有智能化的体现。

其次，目前城市治理中已经开始大规模研究如何结合人工智能技术，也出现了很多示范系平台，但是依旧存在许多尚未解决的问题，主要表现在隐私和应用的效率方面。在隐私方面，由于城市治理相关的数据涉及到个人信息无法公开，阻碍了部门之间数据的流通共享。海量数据可以提供更好的服务，但又会造成隐私泄露的严重后果，且庞大的数据量加剧了部门间资源请求流程的繁琐程度，造成了“信息壁垒”。各部门之间未建立数据协同共享机制，致使海量的政务数据、社会数据被分散在独立的部门或单位之间，形成“信息孤岛、数据鸿沟”现象，数据融合和信息共享不够充分[16]。效率方面，大多城市治理的方案数据来源是手动标注之后再训练模型，导致过高的时间和人力成本，标注数据的质量也会直接影响到模型的效果，因此无监督或者自监督的方法应用就显得尤为重要。此外，由于很多模型的参数过大，造成模型训练和预测时间过长，并不适合实际应用，因此需要在保证准确率的基础上进行优化。

3 接诉即办智能派单业务研究方案

对于接诉即办业务，为达到智能派单效果，本文提出了研究路线，如图 1 所示。对投诉数据进行处理，然后分别构建地址识别基线模型和工单分类模型，为使得分类准确，提出了基于类别比例的标签融合方法，最终对模型结果进行对比分析，并构建投诉回复模板。下面介绍每一步的具体实现过程。

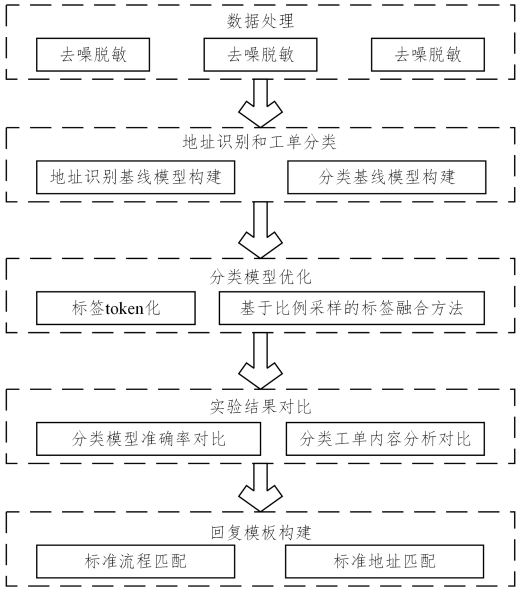

图 1 接诉即办智能派单研究路线

Fig. 1 Intelligent ordering research route

3.1 数据处理

本文中接诉即办智能派单数据主要来自项目收集的北京市海淀区12345热线的咨询记录、人民政府网站上的意见收集(后文称为内部数据)和海淀区领导留言板上公开的数据(后文称为公开数据)。

3.1.1 去噪脱敏

有些数据中存在噪声，如某些意义不明的符号等，其对文本的语义表征作用甚微；此外，某些数据中含有市民隐私信息，如电话、邮箱、身份证号等，故在使用过程中需要进行脱敏工作，隐去个人隐私信息和不能公开的关键信息。

对于这些信息，本文通过构建正则匹配的规则，根据正则表达式对所有文本中的这几种信息进行匹配。在噪声方面，经过统计发现文本中存在连续的“=”、换行符、制表符以及部分工单末尾带有的“注：请及时向来电人反馈办理情况”等噪声数据，使用文本替换的方式将这些噪声去除。在脱敏方面，对于电话、邮箱等信息，首先采用字符串正则匹配的方法将所有含有这些信息的工单抽取出来，利用匹配规则将这些数据进行脱敏。对于手机号码，屏蔽手机号码中间4位数字为“*”号；对于身份证号，屏蔽身份证号中间8位数为“*”号；对于邮箱，只保留前4个字符和邮箱后缀。

3.1.2 标签构建

标签构建的目的是对工单分类，对现有数据添加类别标签。构建标签时，采用《海淀区接诉即办案件处置标准化手册》(后续简称《手册》)作为数据标签的标准，本文的数据对应《手册》中的“城市管理部件”“城市管理事件”和“社会管理服务”3个大类，然后对每个数据类别进行层级细分，使用第三级分类作为标签。初步构建完成后，发现在不同大类中的最低层级标签有重复的现象，如在城市管理事件和社会管理服务中都存在“公园管理”这个子类别，因此单纯使用第三级别的标签无法进行很好的区分，也没有体现出层级结构带有的信息。因此，本文采用拼接的方式将第一级分类与第三级分类拼接到一起，构建成不会出现重复的标签，例如“事件-公园管理”和“服务-公园管理”便于在后续业务中需要添加新数据时不会造成影响，便于扩展。

每一个标签类别对应一套后续处理流程，如处置时限、响应时限、回复标准、参考法律法规等。《手册》中城市管理事件大类的公园管理类别如表1所列，其余类别也都有与表1中公园管理类别相同的标准流程。

表1 接诉即办案件处置标准化手册内容示例

Table 1 Example of standardized manual

一级分类	城市管理事件
二级分类	市容环境
三级分类	公园管理
具体情况	公园内绿地垃圾、附属设施无人管理
主责单位	区园林绿化局
具体操作	24小时核实权属，由权属单位协调，转办处理部门
处置流程	接单后，应仔细查看工单反映人姓名、联系电话、违规行为具体表现以及对周边环境的影响。工单升级后，要立即电话联系或派单至属地街道、镇，及时处理并反馈处理结果。
响应时限	24小时
处理时限	24小时核实
回复标准	已核实处理
法规依据	《关于研究“接诉即办”派单和应急值守工作交接等事宜的会议纪要》

通过对《手册》内容进行规范化抽取，构建出一套完善的三级标签，得到3个大类下共计249个标签类，后续数据标注的标签对应上构建好的249个三级标签类别，每个类别有着对应的标准处理流程。

3.1.3 标签对齐

因为内部数据中存在部分已标注历史数据，其标签名称和上节构建的标签略有差异，但本质是一个类别，例如历史数据中标注的“保险服务”实际对应三级标签中的“保险窗口服务”，因此需要对这些数据进行修正，保持标签的对齐。对于历史数据和新的数据，统计其中的标签类别和占比，将没有直接对应上的标签类别提取出来，将变化不大的类别修正为新标签，对模糊的类别进行讨论决定是否修正或者放弃。经过统计分析，数据的详细情况如表2所列。

表2 接诉即办智能派单数据统计

Table 2 Dataset of intelligent ordering

类别统计	城市管理部件	城市管理事件	社会管理服务
标准二级分类	5	6	9
标准三级分类	112	83	54
标注二级分类	6	6	9
标注三级分类	83	77	107
标注数据量	32767	105138	676922
匹配标签个数	56	49	35
清洗后匹配个数	67	68	93
清洗数据量	17960	86468	641386

表2中，对于3个大类别，其中标准二级分类和标准三级分类表示根据《手册》构建好的标准类别数目，标注二级分类和标注三级分类表示历史数据中标注的统计量，可以看到直接能够匹配上的三级分类的个数都较少，占比为56.2%，经过标签对齐之后，能够匹配上的标签占比为91.6%，最终能使用的数据量约为75万。

标注标签与标准标签对齐的部分示例如表3所列。其中每个大类中左边列为标注标签，右边列为对齐后的标准标签。可以看出某些映射较为清晰，能够直接对应上，但仍旧存在模糊概念，需要进行人为判断。

表3 标准手册标签对齐示例

Table 3 Example of label alignment

城市管理部件		城市管理事件		社会管理服务	
增设过街天桥	过街天桥	群租房	政策性住房	油烟扰民	油烟污染
其他护栏	其他护栏	公共安全	维稳	僵尸车长期占道	停车管理
垃圾间(楼)	垃圾箱、垃圾间	涉法涉诉	违法违纪	商业噪音	噪音扰民
垃圾箱		投诉举报		社会生活噪音	

对于公开数据，通过标签对齐可以将所有数据对应上接诉即办标准手册中经过清洗构建好的标签，数据量为5万左右，经过对齐后将其合并到内部数据中，一共得到约80万条数据。

3.2 基线模型构建

若要实现工单的智能派单，则需要提取投诉工单中的两类重要信息，即工单中提到的地址和工单投诉的事件分类，根据地址能将工单派发到具体部门，根据类别能够映射到标准处理流程库确定处理流程，因此本文需要构建一个地址识别模型和一个工单分类模型，以分别解决这两个问题。

3.2.1 地址识别基线模型构建

在地址识别任务中，本文构建的基线模型是在人民日报1998命名实体识别数据集、人民日报2004命名实体识别数据集和MSRA微软亚洲研究院开源命名实体识别数据集上进行训练的，采用的是基于全词掩码的中文BERT模型[17]，同时也尝试了基础的中文BERT模型[18]、spanBERT模型[19]和复旦大学提出的FLAT模型[20]，最终选择使用基于全词掩码的中文BERT，其在所有公开数据上能达到约92.1%的准确率。在此基础上，本文参考公开数据集的数据格式，通过匹配抽取构建了一份测试数据进行测试，验证模型的准确率，结果如表4所列。

表4 地址识别基线模型的效果

Table 4 Results of baseline of address recognition

	MSRA数据集	PeopleDairy 1998	PeopleDairy 2004	派单数据
数据量	48442	27818	286268	30000
类别数量	3	3	4	2
类别名称	LOC,ORG,PER	LOC,ORG,PER	LOC,ORG,PER,T	LOC,ORG
数据格式	Text:'冬日鼓浪屿,温暖惬意' Labels:[O O B_LOC I_LOC I_LOC O O O O O O]			
准确率	0.95	0.98	0.91	0.93

表4中类别名称表示数据集中标注的类别，其中LOC代表地址，ORG代表组织或建筑的名称，PER表示人名，T表示时间。数据的格式经过预处理后如表4所列，一条数据有text和labels两个部分，其中text表示原文本，labels表示原文本一一对应的数据标签。O代表不属于任何一个类，B_LOC表示该字符是地址开始的字段，I_LOC表示对应这个词的非开始部分，其他类别也进行了同样的处理。由于本文中使用的模型的主要任务为地址识别，因此需要让模型能够识别出地址和建筑物信息，本文只采用其中的LOC和ORG两类。经过实验基线模型能够达到93%的准确率，满足本文业务的需要，因此后续不再对地址识别模型进行进一步的改进，而采用上述训练好的地址识别模型。

3.2.2 工单分类基线模型构建

在分类任务中，本文首先使用常用的BERT系列的预训练模型构建基线模型，经过实验同样采用基于全词掩码的中文BERT模型进行分类，模型结构如图2所示。将文本输入模型进行编码，然后将标签编码为一个对应的编号，将输出的向量经过softmax概率计算，得到准确率最高的标签作为分类结果，使用交叉熵损失作为损失函数优化模型，在城市管理事件、城市管理部件、社会管理服务3个分类的数据集上分别获得了78%，79%和63%的准确率。

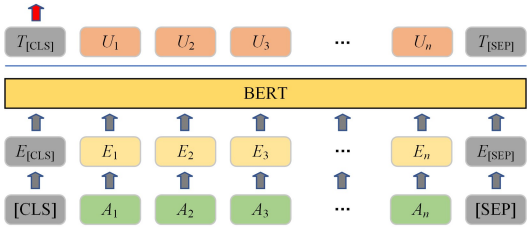

图2 基于BERT的派单分类基线模型

Fig. 2 Baseline of classification based on BERT

分类基线模型中的方法比较依赖数据的有效性，如果数据拥有高噪声或分布不平衡，模型将不能较好地拟合数据分布，从而造成准确率低下的问题。在智能派单业务中，数据集的类别分布也存在高度的不平衡性，为了改善数据类不平衡的影响，本文对基线模型进行如下改进。

首先尝试将训练损失函数改成Focal-loss和CB-loss分别进行训练，Focal-loss通过减少易分类样本的权重，让模型关注于难分类的样本，在平衡交叉熵损失的基础上，增加了一个调节因子，其定义如下：

$$FL(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

其中，$(1-p_t)^{\gamma}$是调节因子，用于减少易分类样本的权重，$\gamma\geqslant 0$是可调节的聚焦参数。

CB-loss中假设一个新采样的数据点只能以概率p从之前采样的数据中获取，或者以概率$(1-p)$在原来的数据集之外采集，从而第n个实例采样后的期望为：

$$E_n=pE_{n-1}+(1-p)(E_{n-1}+1)=1+\frac{N-1}{N}E_{n-1} \tag{2}$$

这里的N表示特征空间中所有数据集合的体积，此时：

$$E_{n-1}=(1-\beta^{n-1})/(1-\beta) \tag{3}$$

$$E_n=(1-\beta^{n})/1-\beta \tag{4}$$

相当于引入一个加权因子来平衡样本总数与有效样本数，将其与其他常见损失函数如交叉熵损失相结合，得到：

$$CB_{CE}(z,y)=-\frac{1-\beta}{1-\beta^{n_y}}\log\left(\frac{\exp(z_y)}{\sum_{j=1}^{C}\exp(z_j)}\right) \tag{5}$$

其中，C表示类别总数，y表示一个样本的标签为y，该标签的训练样本数为n_y。

将基线模型中的交叉熵损失更换为上述两个损失函数，使用Focal-loss函数，基于多分类交叉熵的CB-loss函数和基于Focal-loss的CB-loss函数分别训练模型，测试得到最终的准确率比基线模型只提高了最高1%，依旧没有解决模型准确率低下的问题。

经过查阅文献和参考其他场景下的类不平衡解决方案，本文尝试采用下采样方法对每个类别的数量进行手动控制，设置阈值为3000，5000和10000，超过阈值的类别就随机采样阈值处的样本数进行训练，大幅降低了数据的类不平衡性，得到的结果如表5所列。其中可看到，由于某些类别数据量过大，例如从数十万条中采样3000条，不能很好地拟合原数据分布，导致准确率与基线模型相比有大幅下降。

表5 下采样方法的准确率

Table 5 Accuracy of under-sampling

F1分数	城市管理部件	城市管理事件	社会管理服务
BERT模型	0.7907	0.7985	0.6388
BERT模型+Focal-Loss	0.8024	0.8069	0.6536
BERT模型+CB-Loss	0.7962	0.8044	0.6542
下采样3000	0.2580	0.2304	0.1308
下采样5000	0.2733	0.2717	0.1829
下采样10000	0.2871	0.2935	0.2176

最终，本文决定使用Focal-Loss训练好的BERT模型作为工单分类的基础模型，并在后续基于此模型进行进一步的优化。

3.3 分类模型优化

在3.2节中进行类不平衡处理时可以发现采样的方法效果不佳，因此本节将从标签融合的角度对工单分类模型进行优化。

3.3.1 标签token化

参考文献[21]提出的方法，将标签信息融入finetune

阶段，让文本和标签信息之间的隐含关系能够被学习到，将所有标签的信息进行分词后拼接到文本前面，例如将“非机动车乱停，停车管理，增设停车位，市容市貌”拼接到“市民反映共享单车乱停放问题工单”前面，并且将后者分类到“非机动车乱停”类别，那么模型可以学习到“非机动车乱停”和“市民反映共享单车乱停放问题”之间的隐含关系，同时能学习到与其他类别不相关的知识，从而提高模型的分类能力。

本文首先直接参考文献[21]中提到的方案，将所有的标签拼接到文本前面，一共 249 个标签，然而标签长度已经超过 BERT 模型的最长处理范围 512，同时标签长度过长也会导致模型忽略文本中的知识。

本文决定从预训练模型的词表入手，使用以单个词为编码进行表征的 WoBERT 模型[22]，然后将每一个类别表示为一个 token 加入词表中，同时在分词预处理时增大这些 token 的切分权重，让 249 个标签从原来以字符为单位的多于 800 的长度转换为由 249 个 token 表示，长度固定在 249。接着，文献[23]通过优化模型矩阵参数，来获得更好的拟合效果，经过训练和测试之后，发现该方法依然表现不佳。经分析验证，其原因是将每一个标签用一个 token 表征后，标签在很大程度上丧失了蕴含的语义信息。如对于“垃圾分类”，按照原分词方式，其将由 4 个 token 组成，包含这几个 token 表达的信息，而将其用一个 token 表示后，丧失了这些信息的存储。同时，文本含有的信息往往较为杂乱，因此会导致模型无法较好地收敛。

经过上述实验后，本文决定对城市管理事件、城市管理部件、社会管理服务 3 个类别分开进行训练。在训练时将 3 个部分的标签信息进行拼接，最后得到的效果较差，主要原因是对于每一个大类，都有几十个标签，长度超过 200。而每一个大类别中的工单平均长度并不到 150，因此模型会更多地学习所有的标签信息，而忽略工单文本本身。同时，经过实验发现标签顺序打乱的拼接会比所有文本都依照顺序拼接的准确率高，原因可能为模型的泛化性能得到提升，避免学习到由于顺序带来的隐含信息。考虑到类别的不平衡问题，某些标签中的样本数极少，即使将该标签加入也不能帮助模型学习到文本和标签之间的关系，因此本文提出基于比例采样的标签融合方法。

3.3.2 基于比例采样的标签融合方法

通过分析，本文决定在训练模型时手动预处理数据，将数据的正确标签和随机采样其他 9 个标签拼接到原始文本中进行训练，让模型学习到的不是基线模型中那种文本与一个数字标签之间的关系，而是学习到文本与实际标签内容的关系。在测试阶段，将关键词匹配的标签结果的类别采样比例设置为原类别比例加上一个奖励系数 α，再加入按照标签比例进行采样得到的标签，在文本预处理时，每次采样 10 个标签加入，对于类别极不平衡的社会管理服务部分的数据，对标签的采样概率加入一个惩罚系数 β，将比例超过 10% 的标签采样概率乘上惩罚系数，本文中奖励系数取值为 0.5，惩罚系数为 2。该方法的思想与加入伪标签思想类似，都是通过给模型加入更多的信息，来学习到隐藏的关联关系。然而，本文所采取的方法更为简单，不需要更多的模型来先进行分类再构建伪标签。改进的模型如图 3 所示，实验证明该方法的效果明显。

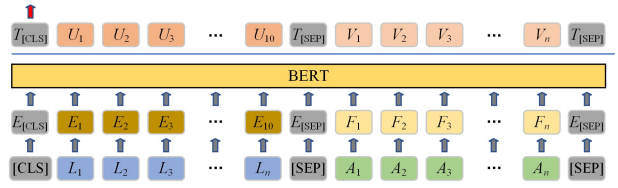

图 3 基于比例采样的标签融合方法

Fig. 3 Label fusion method based on proportional sampling

与基线模型相比，主要是在文本的基础上拼接了随机采样出的 9 个标签和文本对应的真实标签 L1－L10，每一次加入的标签都不一样，能在一定程度上提高模型的泛化能力。

3.4 回复模板构建

本节主要介绍智能派单回复模板的构建，基本流程如图 4 所示，通过回复模板，对输入的工单文本给出回复建议。主要步骤为，首先输入投诉建议或工单内容，调用分类模型对工单内容进行分类，然后检查分类的类别标签对应的《手册》中后续流程的主责单位。接着查询流程中的主责单位部分是否涉及社区、小区和居委会的部分，若存在，则进一步调用地址识别模型抽取工单中的地址信息，将匹配结果所在社区、小区名称构建模板，如“学院路街道办事处”，并与其他后续处理流程一起返回。若检查发现主责单位不涉及这三者，则直接根据分类模型的分类结果进行派单流程的调取。

图 4 中标准地址来源于项目组内部数据中拥有海淀区的房屋建筑和地址的全量信息，根据这些信息构建标准地址库用于地址映射，然后将地址识别结果与标准地址库中的结果进行匹配召回。本文采用的是 BM25 文本匹配算法[24]，经过匹配后获取匹配结果中的标准地址。

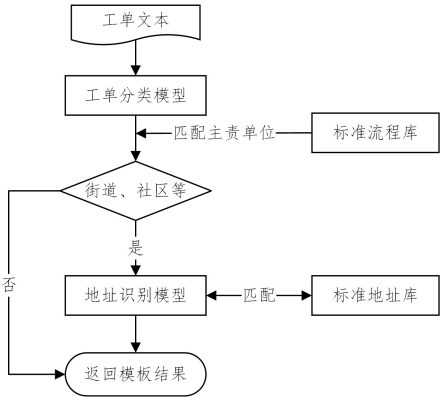

图 4 智能派单回复流程

Fig. 4 Reply process of intelligent order dispatching

4 实验结果及讨论

本节将对智能派单的实验结果进行展示和讨论，其中也对具体工单分类和回复模板构建的示例进行展示，以说明本文采用的方法相比既有基线模型的优势。

4.1 智能派单实验结果

3.2 节阐述了本文对智能派单分类算法所做的一系列尝试，经过一系列实验，所有尝试方案在城市管理部件、城市管理事件和社会管理服务 3 个大类数据集上的结果如表 6 所列。

表 6 智能派单方法结果

Table 6 Results of intelligent order dispatching

准确率	城市管理部件	城市管理事件	社会管理服务
基线模型	0.7907	0.7985	0.6388
+Focal-Loss	0.8051	0.8099	0.6549
+CB-Loss	0.8082	0.8108	0.6587
下采样 3000	0.2580	0.2304	0.1308
对比学习方法	0.3031	0.2780	0.1578
标签融合方法	**0.9103**	**0.8872**	**0.8624**
基线模型 top-3	0.8022	0.8167	0.6530
+Focal-Loss top-3	0.8087	0.8148	0.6622
+CB-Loss top-3	0.8034	0.8189	0.6619
标签融合 top-3	**0.9568**	**0.9322**	**0.8945**

表 6 中准确率指，在仅考虑模型输出的概率最高标签时，计算 F1 分数，在考虑模型输出的 top-3 的 3 个标签时，计算 3 个结果中是否有分类正确的，计算正确样本数占比。可以看到经过标签融合的方法不论是单个标签的准确率还是 top-3 标签的准确率都要比基线模型和其他方法的表现更优秀，尤其是在数据不平衡程度更高的社会管理服务的数据集上，实验结果说明本文所提出标签融合方法的有效性。

实验结果表明，模型的准确率在城市管理事件、城市管理部件和社会管理服务 3 个大类上比基线模型的准确率分别提高了 20%-35%的准确率。本文最后采用的基于类别比例的标签融合方法达到了预期目标。

进一步，本文对部分数据在基线模型上的分类结果和在基于比例采样的标签融合方法的分类结果做了一些对比，表 7 列出了两个工单的示例。两张工单分别反映了房屋交易问题和非机动车乱停放问题，从中可以看到本文方法在一些比较模糊的场景下能够有较高的准确性，而基线模型方法很容易关注工单文本的某个词，从而将工单误判为和此词关系密切的类别。

表 7 部分工单分类结果对比

Table 7 Comparison of work order classification

工单内容	标注	基线	优化
我是海淀区学院路街道成府路 20 号院 50 号楼住户，该楼是新建保障性住房，计划 2019 年 7 月竣工。目前，我们的房子房款已付清、已完成各项验收指标，并交了物业费拿了钥匙，完全符合入住条件。但北京景泰文宇投资有限公司不仅没有建筑资格，在中建一局将房子交付使用后，聚众阻挠 500 余户居民入住，侵犯居民正当、合法诉求与权益，所作所为严重影响社会稳定，性质恶劣，社会影响负能量。	房屋交易	物业管理	房屋交易
市民反映，海淀区海淀医院地铁站 C 口，都是共享单车乱停乱放，很堵，来电反映共享单车乱停放问题。	非机动车乱停	停车管理	非机动车乱停

4.2 回复模板构建结果

对于表 7 中的两个工单文本示例，经过回复模板构建后得到的结果如表 8 所列。表 8 中第一条工单，分类模型将其分到了“房屋交易”类别，该类别在《手册》中对应的主责单位为“属地街道、区房管局”，再通过前文训练好的地址识别模型，抽取出文本中所含地址信息为“海淀区学院路街道成府路 20 号院 50 号楼”，经过地址匹配，对应标准地址库中的结果为“学院路街道-地大第一社区-海淀区成府路 20 号院”，模板回复结果为“海淀区房管局、学院路街道办事处”。表 8 中第二条工单，分类模型将其分到了“非机动车乱停放”类别，该类别在手册中对应的主责单位为“属地街道”，通过前文训练好的地址识别模型，抽取出文本中地址信息为“海淀区海淀医院地铁站 C 口”，通过地址匹配，对应标准地址库中的结果为“海淀街道-海淀南路北社区-海淀医院住院楼-海淀区中关村大街 29 号院”，因此将工单派给“海淀街道办事处”进行处理。

表 8 部分工单派单流程结果

Table 8 Examples of work order dispatching process

工单内容	派单部门
我是海淀区学院路街道成府路 20 号院 50 号楼住户，该楼是新建保障性住房，计划 2019 年 7 月竣工。目前，我们的房子房款已付清、已完成各项验收指标，并交了物业费拿了钥匙，完全符合入住条件。但北京景泰文宇投资有限公司不仅没有建筑资格，在中建一局将房子交付使用后，聚众阻挠 500 余户居民入住，侵犯居民正当、合法诉求与权益，所作所为严重影响社会稳定，性质恶劣，社会影响负能量。	海淀区房管局
市民反映，海淀区海淀医院地铁站 C 口，都是共享单车乱停乱放，很堵，来电反映共享单车乱停放问题。	海淀街道办事处

5 总结与展望

本文主要研究了北京市海淀区接诉即办智能派单业务，通过对该业务的需求进行分析之后，进行数据处理、算法优化、模型测试等相关工作，提出了能够大幅提升该业务效率的一套解决方案。

本文首先进行数据的处理，包括去噪脱敏、标签构建和标签对齐 3 个方面。接下来分别构建了基于公开数据集训练和基于实际数据集作测试的地址信息抽取基线模型和基于 BERT 的派单分类基线模型，并在本文构建好的 3 个派单大类数据集上进行实验，然后分别尝试了基于类不平衡的损失函数、基于采样的方法、基于对比学习的有监督方法，发现在数据集上的提升效果不够明显，甚至有下降。最后参考有研究学者提出的融合标签信息的方法并对其进行了改进，分别对原始方法、Prompt 方法和标签 token 化进行了测试，最后提出基于比例采样的标签融合方法，并在数据集上进行了实验，得到了比基线模型高 10%～25%的提升效果。

但是，本文方法也存在一定的局限性。

1)本文使用的基础模型都是各大科技公司根据公开预训练数据训练好的模型，因此一定程度上模型学习到的参数分布不一定适配城市治理场景下的数据分布，在下游任务上的精调可能会导致模型强行学习城市治理场景下的数据分布而丢失原来学习到的预训练知识。在后续研究中可以考虑在硬件设备支持的情况下收集足够多城市治理场景数据重新训练模型，或研究更深层次的模型泛化，如蒸馏等手段。

2)本文中对于地址匹配只对应了标准地址库中的地址，实际中有一些地址数据无法精准匹配上，后续研究可以对此部分进行改进，提高地址匹配的精度，让此方案能更适配实际场景。

3)本文在数据的预处理上还不够深入，例如在使用 Prompt 时只是尝试了部分论文中提到的模板，很可能该模板并不适配本文的数据集，但是该方法在大量实验中被证明有效，因此后续可以从这方面入手，进一步深挖适用于本文数据的 Prompt 模板。

结束语 接诉即办智能派单是城市治理中的核心业务流程，能加快政务服务的效率和准确性。本文针对派单算法进行优化，提出了基于比例的标签融合方法，给出了一套较为优秀的解决方案。同时，在此基础上对该算法未来的可优化方向进行了展望。

参 考 文 献

[1] XIONG H,YAN H L. Research on the implementation mechanism of data driven delivery platform intelligent dispatch [J]. Nankai Business Management,2022,25(2):15-25.

[2] WU Z C,HONG T,SHEN Z M,et al. Research on intelligent dispatching method for communication defects under service response time constant [J]. Electric Power Information and Communication Technology,2022,20(4):102-107.

[3] LIU J. Research on bank intelligent dispatching system based on improved word embedding representation and convolution neural network [J]. Industrial Control Computer,2020,33(4):101-104.

[4] CHEN P,LI H. Research on O2O takeaway delivery route optimization based on time satisfaction [J]. Chinese Journal of Management Science,2016,24(S1):170-176.

[5] XU Q,XIONG J,YANG Z H,et al. Vehicle routing optimization for takeaway delivery based on adaptive large neighborhood search algorithm [J]. Industrial Engineering and Management,2021,26(3):115-122.

[6] ZHANG L Y,ZHANG J,XIAO B. Research on multi-objective O2O takeaway instant delivery route optimization considering customer priority [J]. Industrial Engineering and Management,2021,26(2),196-204.

[7] LI T Y,LV X N,LI F,et al. Optimization model and algorithm of delivery route considering dynamic demand [J]. Control and Decision,2019,34(2):406-416.

[8] OZAKAN E,WARD A R. Dynamic matching for real-time ride sharing [J]. Stochastic Systems,2020,10(1):1-97.

[9] BRAVERMAN A,DAI J G,LIU X,et al. Empty-car routing in ridesharing systems [J]. Operations Research,2019,67(5).

[10] XU Z,LI Z X,GUAN Q W,et al. Large-scale order dispatch in on-demand ride-hailing platforms:a learning and planning approach [J]. SIGKDD explorations,2018(Udisk):900-908.

[11] AKBARPOUR M,LI S W,GHARAN S O. Thickness and information in dynamic matching markets [J]. Journal of Political Economy,2019,128(3):783-815.

[12] One network unified management:the 'Shanghai plan' for the modernization of mega-city governance [OL]. http://qikan.cqvip.com/Qikan/Article/Detail?id=7103433364.

[13] ZHENG Y. Smart city operating system [J]. Communication of the CCF,2020(2):39-44.

[14] WENG L E. Deepen the reform of 'running at most once' and build an integrated government service model [J]. Chinese Public Administration,2019(6):2.

[15] Notice of the Beijing big data work promotion on printing and distributing the Beijing smart city development action outline during the 'Fourteenth Five-Year Plan' period(JDDF [2001]. No. 1) [OL]. https://www.beijing.gov.cn/zhengce/zhengcefagui/202103/t20210323_2317136.html.

[16] HUANG J. Main problems and suggestions of smart city construction in Fujian Province [J]. Straits Science,2022(2):95-96.

[17] CUI Y M,CHE W X,LIU T,et al. Pre-training with whole word masking for Chinese BERT [J]. arXiv:1906.08101,2019.

[18] WANG Z N,JIANG M,GAO J L,et al. Chinese named entity recognition method based on BERT [J]. Computer Science,2019,46(S2):138-142.

[19] JOSHI M,CHEN D Q,LIU Y H,et al. SpanBERT:improving pre-training by-representing and predicting spans [J]. Transactions of the Association for Computational Linguistics,2020(8):64-77.

[20] LI X N,YAN H,QIU X P,et al. FLAT:Chinese NER using flat-lattice transformer [J]. arXiv:2004.11795,2020.

[21] XIONG Y J,FENG Y K,WU H,et al. Fusing label embedding into BERT:an efficient improvement for text classification[C]// Proc of Findings of the Association for Computational Linguistics:ACL-IJCNLP2021,2021. 2021:1743-1750.

[22] SU J L. WoBERT:word-based Chinese BERT model-ZhuiyiAI [EB/OL]. (2020-09-18)[2023-09-16]. https://kexue.fm/archives/7758.

[23] WU X,GAO C C,LIN M,et al. Text smoothing enhance various data augmentation methods on text classification tasks [J]. arXiv:2202.13840,2022.

[24] ROBERTSON S,ZARAGOZA H. The probabilistic relevance framework:BM25 and beyond [J]. Foundations & Trends in Information Retrieval,2009,3(4):333-389.

JIA Jingdong, born in 1975, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. Her main research interests include natural language processing, software testing, machine learning and requirements engineering.