软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.005977]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

数据标注研究综述*

蔡 莉[1,2*], 王淑婷[1], 刘俊晖[1], 朱扬勇[2]

[1](云南大学软件学院,云南 昆明,650091)

[2](复旦大学计算机科学技术学院,上海,200433)

通讯作者: 蔡莉, E-mail: caili@ynu.edu.cn

摘 要: 数据标注是大部分人工智能算法得以有效运行的关键环节.数据标注越准确、标注的数据量越大,算法的性能就越好.数据标注行业的发展带动了中国许多城市和城镇的就业,促使中国逐渐成为世界数据标注的中心.本文阐述了数据标注的发展概况,包括起源、应用场景、分类和任务;列举了目前常用的标注数据集、开源的数据标注工具和商业数据标注平台;提出了标注中的角色、标准和流程等数据标注规范;给出了一个情感分析场景中的数据标注实例;描述各类主流的标注质量评估算法及其特点,并对比它们优缺点;最后,从任务、工具、数据标注质量和安全性四个方面对数据标注的研究方向和发展趋势进行了展望.

关键词: 数据标注;人工智能;众包;大数据

中图法分类号: TP391

中文引用格式: 蔡莉,王淑婷,刘俊晖,朱扬勇.数据标注研究综述.软件学报. http://www.jos.org.cn/1000-9825/5977.htm

英文引用格式: Cai L, Wang ST, Liu JH, Zhu YY. Behind Artificial Intelligence: A Survey of Data Annotation. Ruan Jian Xue Bao/Journal of Software, (in Chinese). http://www.jos.org.cn/1000-9825/5977.htm

Behind Artificial Intelligence: A Survey of Data Annotation

CAI Li[1,2], WANG Shu-Ting[1], LIU Jun-Hui[1], ZHU Yang-Yong[2]

[1](School of Software, Yunnan University, Kunming 650091, China)

[2](School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: Data annotation is a key part of the effective operation of most artificial intelligence algorithms. The better the annotation accuracy and quantity, the better the performance of the algorithm. The development of the data annotation industry boosts employment in many cities and towns in China, prompting China to gradually become the center of world data annotation. This paper summarizes its development, including origin, application scenarios, classifications, and tasks; lists the commonly used annotation data sets, open source data annotation tools and commercial annotation platforms; proposes the data annotation specification including roles, standards, and processes; gives an example of data annotation in a sentiment analysis. Then, this paper describes the models and characteristics of state of art algorithms for evaluating annotation results, and compares their advantages and disadvantages. Finally, this paper prospects research focuses and development trends of data annotation from four aspects: tasks, tools, annotation quality and security.

Key words: Data Annotation; Artificial Intelligence; Crowdsourcing; Big Data

近年来,作为人工智能（Artificial Intelligence,AI）的核心技术,深度学习在图像、语音、文本处理等领域取得了大量关键性突破.尤其在 2016 年和 2017 年,由 Google 公司开发的 AlphaGo 围棋机器人利用深度学习技

* 基金项目: 国家自然科学基金(61663047, U1636207); 云南大学服务云南行动计划项目(2016ZD05)

Foundation item: National Natural Science Foundation of China (61663047, U1636207); The Project of Servicing Yunnan by YNU (Yunnan University) (2016ZD05)

收稿时间: 2019-06-22; 修改时间: 2019-08-05,2019-09-17; 采用时间: 2019-10-30; jos 在线出版时间: 2019-12-05

CNKI 网络优先出版: 2019-12-05 14:55:16, http://kns.cnki.net/kcms/detail/11.2560.TP.20191205.1454.008.html

术完善了围棋算法,分别战胜围棋界的世界冠军李世石和柯洁,震惊了整个科技界[1].人工智能是机器产生的智能,在计算机领域是指根据对环境的感知,做出合理的行动并获得最大收益的计算机程序[2].也就是说,要想实现人工智能,需要把人类理解和判断事物的能力教给计算机,让计算机拥有类似人类的识别能力[3].人类在认识一个新事物时,首先要形成对该事物的初步印象,例如,要识别出飞机就需要看到相应的图片或者真实物体.数据标注可视为模仿人类学习过程中的经验学习,相当于人类从书本中获取已有知识的认知行为.具体操作时,数据标注把需要计算机识别和分辨的图片事先打上标签,让计算机不断地识别这些图片的特征,最终实现计算机能够自主识别[4].数据标注为人工智能企业提供了大量带标签的数据供机器训练和学习,保证了算法模型的有效性.

1 数据标注概述

1.1 数据标注的起源

2007 年,斯坦福大学教授李飞飞等人开始启动 ImageNet 项目,该项目主要借助亚马逊的劳务众包平台 Mechanical Turk（AMT）来完成图片的分类和标注,以便为机器学习算法提供更好的数据集[5].截至到 2010 年,已有来自 167 个国家的 4 万多名工作者提供了 14,197,122 张标记过的图片,共分成 21,841 种类别[6].从 2010 年到 2017 年,ImageNet 项目每年举办一次大规模的计算机视觉识别挑战赛,各参赛团队通过编写算法来正确分类、检测和定位物体及场景.ImageNet 项目的成功改变了人工智能领域中大众的认知,即数据是人工智能研究的核心,数据比算法重要得多[7].从此,数据标注拉开了序幕.目前,学术界尚未对数据标注的概念形成一个统一的认识,比较认可的是由王翀和李飞飞等人提出的定义.他们认为：标注[8]是对未处理的初级数据,包括语音、图片、文本、视频等进行加工处理,并转换为机器可识别信息的过程.原始数据一般通过数据采集获得,随后的数据标注相当于对数据进行加工,然后输送到人工智能算法和模型里完成调用[9].数据标注产业主要是根据用户或企业的需求,对图像、声音、文字等对象进行不同方式的标注[10],从而为人工智能算法提供大量的训练数据以供机器学习使用[11]. 图 1 显示了一个图像标注的示例,标注者需要识别和标注图片中的景物如天空、树木、建筑、湖水和草等对象.

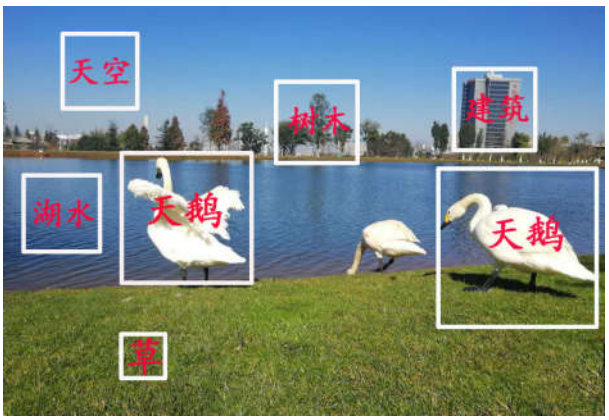

Fig.1 Example of data annotation

图 1 数据标注示例

1.2 数据标注的应用场景

数据标注产业的发展促进了人工智能的蓬勃兴起,其主要的应用行业和不同行业的标注场景总结如下.

(1) 自动驾驶[12]：利用标注数据来训练自动驾驶模型,使其能够感知周围的环境并在很少或没有人为输入的情况下移动.自动驾驶中的数据标注涉及行人识别、车辆识别、红绿灯识别、道路识别等内容,可以为相关

企业提供精确的训练数据,为智能交通保驾护航.

(2) 智能安防[13]：数据标注扩大了现有安防系统的感知范围,通过融合各种来源的数据并进行协同分析,提高监控和报警的准确性;其对应的标注场景有面部识别、人脸探测、视觉搜索、人脸关键信息点提取以及车牌识别等.

(3) 智慧医疗：人工智能和大数据分析技术应用于医疗行业,可以深入洞察医学知识和数据,帮助医生和患者解决在医学影像、新药研发、肿瘤与基因、健康管理等领域所面临的影像识别困难、药物研发成本巨大、癌症治疗效果不佳等难题[14].其所涉及的场景有手术工具标识、处方识别、医疗影像标注、语音标注等.

(4) 工业 4.0：利用标注数据训练和验证机器人应用程序的计算机视觉模型,从而使模型对工业环境内的各类障碍物、机械设备和机器人有更加精确的感知[15],实现工业智能机器与所处环境中人和物的安全交互.对应的场景有机械手臂导航、仓储码垛、自动分拣或抓取、自动焊接等.

(5) 新零售：将人工智能和机器学习应用于新零售行业,可以通过商品销售数据以及用户的真实反馈促进电子商务的销售,提高用户的个性化体验以及预测客户需求[16],并实现线上货物推荐的精准化.新零售中涉及的标注场景包括超市货架识别、无人超市系统和电子商务智能搜索与推荐等.

(6) 智慧农业：依托精准的数据标注实现对农作物的定位以及对其成熟度和生长状态的识别,实现农作物智能采摘并解决精准农药撒播问题[17],从而减少人力消耗并提高农药利用率.目前,智慧农业中有关数据标注的场景有栽培管理、精准水肥和安全监测等.

1.3 数据标注的分类方法

本节详细比较了不同数据标注分类方法的概念和优缺点,如表 1 所示：

Table 1 Classification of data annotation

表 1 数据标注分类

分类方式	分类方法	概念	优点	缺点
标注对象	图像标注	图像标注和视频标注统称为图像标注	使人脸识别和自动驾驶等技术得到发展和完善	相对复杂,耗时
	语音标注	需要人工将语音内容转录为文本内容,然后通过算法模型识别转录后的文本内容	帮助人工智能领域中的语音识别功能更加完善	算法无法直接理解语音内容,需要进行文本转录
	文本标注	与音频标注有些相似,都需要通过人工识别转录成文本形式	减少了文本识别行业和领域的人工工作量	人工识别过程繁杂
标注的构成形式	结构化标注	数据标签必需在规定的标签候选集合内,标注者通过将标注对象与标签候选集合进行匹配,选出最合理的标签值作为标注结果[18]	标签候选集将标注类别描述得很清晰,便于标注者选择;标签是结构化的,利于存储和后期的统计查找[19]	遇到具有二义性标签时往往会影响最终的标注结果
	非结构化标注	标注者在规定约束内,自由组织关键字对标注对象进行描述[20]	给标注者足够的自由,可以清楚地表达自己的观点	给数据存储和使用带来困难,不利于统计分析
	半结构化标注	标签值是结构化标注,而标签域是非结构化标注[21]	标注灵活性强,便于统计查找	对标注者的要求高,工作量高,耗时
标注者类型	人工标注	雇用经过培训的标注员进行标注	标注质量高	标注成本高,时间长,效率低
	机器标注	标注者通常是智能算法	标注速度快,成本相对较低	算法对涉及高层语义的对象识别和提取效果不好

如上表所示,目前数据标注有三种常用的划分方式,(1) 按照标注对象进行分类,包括：图像标注、视频标注、语音标注和文本标注[22].(2) 根据标注的构成形式,将其分为结构化标注、非结构化标注和半结构化标注[23-26].(3) 根据标注者类型,分为人工标注和机器标注[27].

图像标注包括图像标注和视频标注,因为视频也是由连续播放的图像所组成[28].图像标注一般要求标注人员使用不同的颜色来对不同的目标标记物进行轮廓识别,然后给相应的轮廓打上标签,用标签来概述轮廓内的内容,以便让算法模型能够识别图像中的不同标记物[29,30].图像标注常用于人脸识别、自动驾驶车辆识别等应

用[31].语音标注是通过算法模型识别转录后的文本内容并与对应的音频进行逻辑关联[32].语音标注的应用场景包括自然语言处理、实时翻译等,语音标注的常用方法是语音转写.文本标注是指根据一定的标准或准则对文字内容进行诸如分词、语义判断、词性标注、文本翻译、主题事件归纳等注释工作,其应用场景有名片自动识别、证照识别等[33].目前,常用的文本标注任务有情感标注、实体标注、词性标注及其它文本类标注.图 2 以文本标注中的中文文本词性标注为例进行说明,其中,n、v和a分别代表句子中词语的词性,即n表示名词、v表示动词、a表示形容词、wp代表断句.

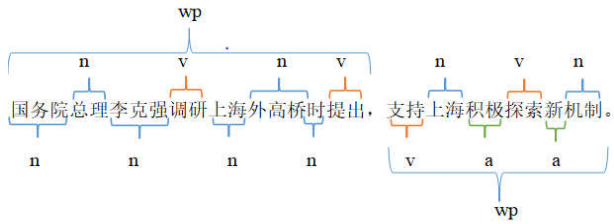

Fig.2 Example of text annotation

图 2 文本标注示例

下文 1.4 小节中提到的标框标注就是典型的半结构化标注,如在豆瓣影评上对某部电影进行评价或在小组会议上发表自己的观点等都属于非结构化标注,而对电影类型进行标注就属于结构化标注.人工标注和机器标注比较好理解,这里就不再举例.除了表 1 所列举的分类方法外,元数据标注也受到了一些学者的关注.郭晓明等人使用基于相似度计算的语义标注算法DM-SAAS(Database Metadata Semantic Annotation Algorithm based on Similarity)实现了元数据自动语义标注[34],也为数据标注工作研究者提供了思路.

1.4 数据标注的任务

常见的数据标注任务包括分类标注、标框标注、区域标注、描点标注和其他标注等.下面介绍每一种任务的具体内容[35].

(1) 分类标注.分类标注是从给定的标签集中选择合适的标签分配给被标注的对象[36].通常,一张图可以有很多分类/标签,如运动、读书、购物、旅行等.对于文字,又可以标注出主语、谓语、宾语,名词和动词等[37].此项任务适用于文本、图像、语音、视频等不同的标注对象[38].本文以图像的分类标注为例进行说明,如图 3 所示.图 3 显示了一张公园的风景图,标注者需要对树木、猴子、围栏等不同对象加以区分和识别.

(2) 标框标注.标框标注就是从图像中选出要检测的对象[39],此方法仅适用于图像标注.标框标注可细分为多边形拉框和四边形拉框两种形式.多边形拉框是将被标注元素的轮廓以多边型的方式勾勒出来,不同的被标注元素有不同的轮廓,除了同样需要添加单级或多级标签以外,多边型标注还有可能会涉及到物体遮挡的逻辑关系从而实现细线条的种类识别[40].四边形拉框主要是用特定软件对图像中需要处理的元素(比如：人、车、动物等)进行一个拉框处理,同时用一个或多个独立的标签来代表一个或多个需要处理的元素.例如,图 4 对人物的帽子进行了多边形拉框标注,图 5 则对天鹅进行了四边形拉框标注.

(3) 区域标注.相比于标框标注,区域标注的要求更加精确[41],而且边缘可以是柔性的,并仅限于图像标注,其主要的应用场景包括自动驾驶中的道路识别和地图识别等.在图 6 中,区域标注的任务是在地图上用曲线将城市中不同行政区域的轮廓形式勾勒出来,并用不同的颜色（浅蓝、浅棕、紫色和粉色）加以区分.

(4) 描点标注.描点标注是指将需要标注的元素（比如人脸、肢体）按照需求位置进行点位标识,从而实现特定部位关键点的识别[42].例如,图 7 采用描点标注的方法,对图示人物的骨骼关节进行了描点标识.人脸识别、骨骼识别等技术中的标注方法与人物骨骼关节点的标注方法相同[43].

(5) 其他标注.数据标注的任务除了上述 4 种以外,还有很多个性化的标注任务.例如,自动摘要就是从新闻事件或者文章中提取出最关键的信息,然后用更加精炼的语言写成摘要[44].自动摘要与分类标注类似,但两者存

在一定差异.常见的分类标注有比较明确的界定,比如在对给定图片中的人物、风景和物体进行分类标注时,标注者一般不会产生歧义.而自动摘要需要先对文章的主要观点进行标注,这相比于分类标注来说,在标注的客观性和准确性上都没有像分类标注那么严格,所以自动摘要不属于分类标注.

Fig.3 Classification annotation
图 3 分类标注

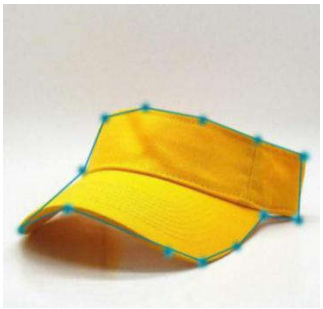

Fig.4 Polygonal frame
图 4 多边形拉框

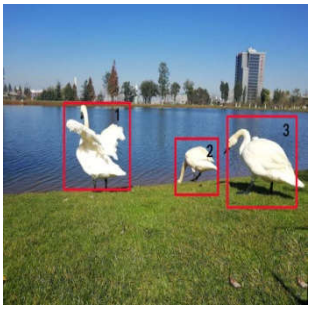

Fig.5 Quadrilateral frame
图 5 四边形拉框

Fig.6 Region annotation
图 6 区域标注

Fig.7 Point annotation
图 7 描点标注

2 数据标注中的数据集、工具和平台

随着人工智能、机器学习等行业对标注数据的海量需求,许多企业和研究机构纷纷推出了带标注的公开数据集.为了提高数据标注效率,一些标注工具和平台也应运而生[45] .下面,将对常用的标注数据集,部分主流的数据标注工具、平台及其适用场合进行阐述.

2.1 常用标注数据集

本文将标注数据集划分为图像、视频、文本和语音标注数据集四大类,表 2 描述了这些数据集的来源、用途和特性.ImageNet、COCO 和 PASCAL VOC 是三个典型的图像标注数据集,它们广泛应用于图像分类、定位和检测的研究中.由于 ImageNet 数据集拥有专门的维护团队,而且文档详细,它几乎成为了目前检验深度学习图像领域算法性能的“标准”数据集.COCO 数据集是在微软公司赞助下生成的数据集,除了图像的类别和位置标注信息外,该数据集还提供图像的语义文本描述.因此,它也成为评价图像语义理解算法性能的“标准”数据集.Youtube-8M 是谷歌公司从 YouTube 上采集到的超大规模的开源视频数据集,这些视频共计 8 百万个,总时长为 50 万小时,包括 4800 个类别.Yelp 数据集由美国最大的点评网站提供,包括了 470 万条用户评价,15 多万条商户信息,20 万张图片和 12 个城市信息.研究者利用 Yelp 数据集不仅能进行自然语言处理和情感分析,还可以用于图片分类和图像挖掘.Librispeech 数据集是目前最大的免费语音识别数据库之一,由近 1000 小时的多人朗读的清晰音频及其对应的文本组成,它是衡量当前语音识别技术最权威的开源数据集.

Table 2 Partial annotation datasets

表 2 部分常用的标注数据集

类别	数据集名称	用途	大小	来源/机构	开放
图像标注数据集	ImageNet	图像分类、定位、检测	~1TB	http://www.image-net.org/about-stats	是
	COCO	图像识别、分割和图像语义	~40G	http://mscoco.org/	是
	PASCAL VOC	图像分类、定位、检测	~2GB	http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html	是
	Open Image	图像分类、定位、检测	~1.5GB	https://storage.googleapis.com/openimages/web/index.html	是
	Flickr30k	图片描述	30MB	http://shannon.cs.illinois.edu/DenotationGraph/data/index.html	是
视频标注数据集	Youtube-8M	理解和识别视频内容	1PB	https://research.google.com/youtube8m/	受限
	Kinetics	动作理解和识别	~1.5TB	https://deepmind.com/research/open-source/open-source-datasets/kinetics/	是
	AVA	人类动作识别	-	https://research.google.com/ava	是
	UCF101	视频分类、动作识别	6.5GB	http://crcv.ucf.edu/papers/UCF101_CRCV-TR-12-01.pdf	是
文本标注数据集	Yelp	文本情感分析	~2.66G	https://www.yelp.com/dataset/challenge	是
	IMDB	文本情感分析	80.2MB	http://ai.stanford.edu/~amaas/data/sentiment/	是
	Multi-Domain Sentiment	文本情感分析	52MB	http://www.cs.jhu.edu/~mdredze/datasets/sentiment/	是
	Sentiment 140	文本情感分析	80MB	http://help.sentiment140.com/	是
语音标注数据集	LibriSpeech	训练声学模型	~60GB	http://www.openslr.org/12/	是
	AudioSet	声学事件检测	80 MB	https://research.google.com/audioset/	是
	FMA	语言识别	~1000G	https://github.com/mdeff/fma	是
	VoxCeleb	语音识别、情绪识别	150MB	http://www.robots.ox.ac.uk/~vgg/data/voxceleb/	是

2.2 商业的数据标注平台

通常,商用的数据标注工具一般是由众包标注平台来提供.数据标注众包模式[46,47]平台最早出现在美国,除了亚马逊的 Mechanical Turk[48]平台外,还有 Figure-eight、CrowdFlower、Mighty AI 等初创型标注平台[49].近年来,国内的一些互联网公司、大数据公司和人工智能公司也纷纷推出了自己的数据标注众包平台和商用标注工具,如数据堂、百度众测、阿里众包、京东微工等.这些商业的数据标注平台基本上都能对图片、视频、文本和语音等数据进行标注,但各自的业务方向也有一定侧重,有的以处理图像见长,有的则更擅长做一些视频标注[50].

无论开源的标注工具还是商用的数据标注平台,它们至少要包含[51]:(1) 进度条:用于指示数据标注的进度.一方面方便标注人员查看进度,另一方面也利于统计;(2) 标注主体(指需要标注的对象):可以根据标注形式进行设计,一般可以分为单个标注(指对某一个对象进行标注)和多个标注(指对多个对象进行标注)的形式[52];(3) 数据导入导出功能;(4) 收藏功能:针对模棱两可的数据,可以减少工作量并提高工作效率;(5) 质检机制:通过随机分发部分已标注过的数据,检测标注人员的可靠性.

2.3 开源的数据标注工具

在选择数据标注工具时,需要考虑标注对象(如图像、视频、文本等)、标注需求(如画框、描点、分类等)和不同的数据集格式[53](比如COCO,PASCAL VOC,JSON等).常用标注工具如下表 3 所示. 上表中列举了一些开源的数据标注工具及其特点.表 3 中除了COCO UI和LabelMe工具在使用时需要MIT许可外,其他的工具均为开源使用.大部分的开源工具都可以运行在Windows、Linux、Mac OS系统上,仅有个别工具是针对特定操作系统开发的(如RectLabel);而且,这些开源工具大多只针对特定对象进行标注,只有少部分工具(如精灵标注助手)能同时标注图像、视频和文本.除了表 3 中列举的标注工具外,市场上还有一些特殊功能的标注工具,例如人脸数据标注[54]和 3D点云标注工具.不同标注工具的标注结果会有一些差异,但尚未有研究关注它们的标注效率和标注结果的质量[55].

Table 3 Partial open source data annotation tools

表 3 部分开源的数据标注工具

名称	简介	运行平台	标注形式	导出数据格式
LabelImg	著名的图像标注工具	Windows、Linux、Mac	矩形	XML 格式
LabelMe	著名的图形界面标注工具,能标注图像和视频	Windows、Linux、Mac	多边形、矩形、圆形、多段线、线段、点	VOC 和 COCO 格式
RectLabel	图像标注	Mac	多边形、矩形、多段线线段、点	YOLO、KITTI、COCO1 与 CSV 格式
VOTT	微软发布的基于 Web 方式本地部署的标注工具,能标注图像和视频	Windows、Linux、Mac	多边形、矩形、点	TFRecord、CSV、VoTT 格式
LabelBox	适用于大型项目的标注工具,基于 Web、能标注图像、视频和文本	--	多边形、矩形、线、点、嵌套分类	JSON 格式
VIA	VGG(Visual Geometry Group)的图像标注工具,也支持视频和音频标注	--	矩形、圆、椭圆、多边形、点和线	JSON 格式
COCO UI	用于标注 COCO 数据集的工具,基于 Web 方式	--	矩形、多边形、点和线	COCO 格式
Vatic	Vatic 是一个带有目标跟踪的视频标注工具,适合目标检测任务	Linux	--	VOC 格式
BRAT	基于 Web 的文本标注工具,主要用于对文本的结构化标注	Linux	--	ANN 格式
DeepDive	处理非结构化文本的标注工具	Linux	--	NLP 格式
Praat	语音标注工具	Windows、Unix、Linux、Mac	--	JSON 格式
精灵标注助手	多功能标注工具	Windows、Linux、Mac	矩形、多边形和曲线	XML 格式

3 数据标注规范

3.1 数据标注的角色

传统手工数据标注中的用户角色可以分为三类[56]：(1) 标注员：负责标注数据,通常由经过一定专业培训的人员来担任.在一些特定场合或者对标注质量要求极高的行业（例如,医疗）也可以直接由模型训练人员（程序员）或者领域专家来担任.(2) 审核员：负责审核已标注的数据,完成数据校对和数据统计,适时修改错误并补充遗漏的标注.这个角色往往由经验丰富的标注人员或权威专家来担任.(3) 管理员：负责管理相关人员,发放和回收标注任务.数据标注过程中的各个角色之间相互制约,各司其职,每个角色都是数据标注工作中不可或缺的一部分.此外,已标注的数据往往用于机器学习和人工智能中的算法,这就需要模型训练人员利用人工标注好的数据训练出算法模型.而产品评估人员则需要反复验证模型的标注效果并对模型是否满足上线目标进行评估.

3.2 数据标注的质量标准

本小节根据标注对象本身的特征和标注需求来阐述数据标注要遵循的质量标准[57],当然,在实际操作中还需要根据实际情况进一步细化.

1. 图像标注的质量标准.

机器学习中图像识别的训练是根据像素点进行的,因此,图像标注的质量好坏取决于像素点的判定准确性.如果标注像素点越接近标注物的边缘像素,则标注质量越高,标注难度就越大;反之,则标注质量较差,标注难度较小.按照 100%准确率的图像标注要求,标注像素点与标注物的边缘像素点的误差应该在 1 个像素以内[59].

2. 语音标注的质量标准

在进行语音标注时,标注员需要时刻关注语音数据发音的时间轴与标注区域的音标是否同步.所以,标注与发音时间轴的误差要控制在 1 个语音帧以内.如果误差超过 1 个语音帧,很容易标注到下一个发音,从而产生更多的噪声数据.

3. 文本标注的质量标准

由于文本标注中的任务较多,不同任务的质量标准各有不同.例如,中文分词的质量标准是标注好的分词与词典中的词语一致,不存在歧义.情感标注的质量标准则要求对标注句子的情感分类级别正确.多音字标注的质量标准是借助专业性工具（如字典）来标注一个字的全部读音;而语义标注的质量标准是标注清楚词语或句子的真实语义.

3.3 数据标注的流程

本小节以众包模式下的数据标注为例,提出了一个完整的数据标注流程,如图 8 所示. 数据标注流程首先从标注数据的采集[58]开始,采集的对象包括视频、图片、音频和文本等多种类型和多种格式的数据.由于采集到的数据可能存在缺失值、噪声数据、重复数据等质量问题,故首先需要执行数据清洗任务[59],以便获得高质量的数据,然后对清洗后的数据进行标注,这是数据标注流程中最重要的一个环节.在具体流程中,管理员会根据不同的标注需求,将待标注的数据划分为不同的标注任务.每一个标注任务都有不同的规范和标注点要求,并且一个标注任务将会分配给多个标注员完成.数据标注员完成标注工作后将相关数据交给模型训练人员,后者利用这些标注好的数据来训练出需要的算法模型.标注数据的质量主要由审核员来检验,审核员进行模型测试并将测试结果反馈给模型训练人员,而模型训练人员通过不断地调整参数,以便获得性能更好的算法模型.如果经过参数调整后不能得到最优的算法模型,则说明已标注的数据不满足需求.这时,审核员就会向标注员反馈数据问题,标注员则要重新标注数据.最后,审核员将最优模型指标发送给产品评估人员使用并进行上线前的最后评估.

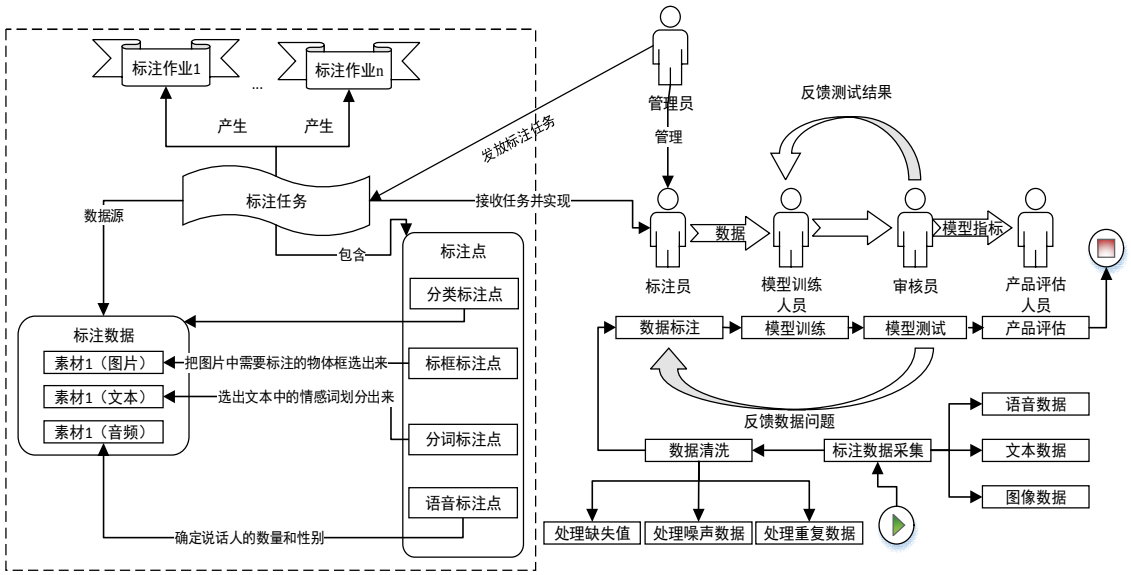

Fig.8 Basic process of data annotation

图 8 数据标注基本流程

4 数据标注实例——情感分析

4.1 情感分析概述

随着电子商务、社交网络和移动互联网的蓬勃发展,互联网上出现了大量带有情感色彩的文本数据.针对文本数据的情感分析能够帮助政府部门及企事业单位更好地理解用户的观点,并及时解决出现的各类问题,以进一步提高服务质量[60].因此,情感分析广泛应用在舆情管控、商业决策、观点搜索、信息预测和情绪管理等场景.

词语、句子和文章是文本情感分析中的三个级别[61].词语级别的情感分析用来确定词语的情感倾向方向和强度,句子级别的情感分析是先对句子进行命名实体识别和句法分析,在采用基于词典和机器学习的研究方法对句子进行情感分析.文章级的情感分析则是分析文章段落的情感倾向方向.情感倾向是主体对某一客体主观存在的内在评价的一种倾向,它由情感倾向方向和情感倾向度来衡量[62],情感倾向方向也称为情感极性.在情绪文本中,情感倾向方向是用户对客体表达其自身观点态度,即支持（正面情感）、反对（负面情感）、中立（中性情感）;;情感倾向度是指主体对客体表达正面情感或负面情感时的强弱程度,不同的情感程度往往是通过不同的情感词或情感语气等来体现.在情感倾向分析研究中,通过对每个情感词赋予不同的权值来区分两者的程度.

4.2 情感分析中的数据标注

情绪文本的分析和挖掘涉及到文本数据标注中的多项任务,下面将对这些任务进行阐述.

1. 中文分词.中文分词是将一个汉字序列切分为一个个单独的词,中文分词是汉语文本处理的基础.例如：要判断句子A=“今天是国庆节,可是我们还要加班.”的情感,首先要将其切分为一个个单词,如果采用自动分词,其结果为：

“今天/是/国庆节/,/可是/我们/还/要/加班/.”

如果采用基于字标注的分词方法,则其结果为：

“今/B天/E是/S国/S庆/M节/E/,/S/可/S是/E我/B们/E还/S要/S加/S班/E./S”

其中,B表示词首,M表示词中,E表示词尾,S代表单独成词,它们形成了四个构词位置.

2. 词性标注.词性标注是将词划分为对应的语法分类,以表达这个词在上下文中的含义.词的语法分类主要为：名词、动词、形容词、量词、代词、副词、连词、助词等.上述句子A的词性标注结果为：

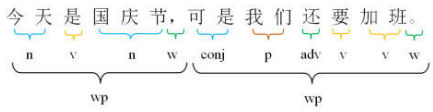

Fig.9 Examples of POS annotation

图 9 词性标注示例

其中,n、v、conj、p和adv分别代表句子中的名词、动词、连词、代词和副词,w表示标点符号,wp代表断句.

3. 情感标注.句子A中并没有明确表示情绪的词,不过联系上下文,可知句子表达的情绪是“低落”.为了判断句子A所表达的情绪,我们可以使用一些中文情感极性词典进行分析,比如来源于台湾大学的 NTUSD 和知网的情感极性字典.但是,本例中如果只依靠中文情感极性词典,计算机很难准确判断句子A所反映的真实情绪.因此,事先要采用人工标注的方法来对一些带情绪的语句进行情感标注.通常,人类的基本情绪可以划分为六种,即快乐、愤怒、悲伤、恐惧、惊讶和嫉妒.为了正确识别情绪,每一类情绪都要有对应的标注数据.然后,利用这些带情绪标注的数据集来训练情绪分类模型.情绪分类算法可以采用K最近邻（K-Nearest Neighbor,KNN）、支持向量机（Support Vector Machines,SVM）、深度置信网络（Deep Belief Network,DBN）和长短期记忆网络（Long Short-Term Memory,LSTM）等实现.一旦分类模型训练成功,就能准确识别句子A所表达的情绪.

5 数据标注质量评估

本文按照数据标注对象,将数据标注结果评估算法分为图像（含视频）、文本和语音三类标注结果评估算法,下面按照时间顺序对这三类评估算法进行简要概述.

5.1 图像标注质量评估算法

目前,比较常用的图像标注质量评估算法[63]有多数投票算法(Majority Voting,MV)、期望最大化算法(Expectation Maximization,EM)和 RY 算法.MV 是由约翰逊提出的一种通用性强的质量控制算法,它将绝大多数用户选择的结果视为最终结果[64-68].MV 算法的基本思想是：假设有 m 个图像标注任务$(t_1,\ldots,t_m)$,每个任务 t_i 对

应一个二元分类,任务管理员将这些任务分配给众包平台中的员工,其中 W 代表所有员工的集合.为了提高标注质量和标注的可靠性,将需要标注的对象 x_i 提供给 N（$N=\{w_1,...,w_N\}\le W, w_j\in W$）个工人进行标注.每个工人 w_j 对 x_i 做出预测并创建一个标签 $y_i^j=w_j(x_i)\in\{0,1\}$,然后根据标签 $\{y_i^1\ldots\ldots y_i^N\}$ 推断出 x_i 的最终标签.其公式如下：

$$\hat{y}_i=\begin{cases}1, & \frac{1}{N}\sum_{j=1}^{N}y_i^j>\frac{1}{2},\\ random\ \text{guess}, & \frac{1}{N}\sum_{j=1}^{N}y_i^j=\frac{1}{2},\\ 0, & \frac{1}{N}\sum_{j=1}^{N}y_i^j=\frac{1}{2}.\end{cases}\tag{1}$$

由于 MV 算法把大多数人认为正确的标签作为最终标签且简单易用,所以常被其他众包质量评估算法当作基准算法.但是,现实生活中大多数人认为正确的并不总是正确,为了解决这个问题,Dawid[69]人提出了 EM 算法.EM 算法[70-72]需要构建出一次标注任务中标注者的标注错误率混淆矩阵,并与实际观测标注结果进行比较,二者比较结果的差异越大,就代表标注的结果越差.RY 算法[73,74]是在 MV 算法和 EM 算法的基础上的改进算法[75,76].Raykar 和 Yu 等人利用 two-coin 模型,在公式(1)的基础上,描述每个数据标注者的素质并估计了敏感性(specificity)、特异性(sensitivity)的概率,通过对标注者的特异性和敏感性进行建模分析,过滤垃圾标注者并提高标注者的质量.

上述三种算法中的 MV 算法和 EM 算法主要用于标注者质量未知的情况下,它们可以检测并剔除低质量的标注者,如果检测到的低质量标注者越多则说明标注质量越差.

5.2 文本标注质量评估算法

常用的文本标注质量评估算法有六种：(1) Kishore Papineni 等人提出的 BLEU(Bilingual Evaluation understudy)[77]算法是一种基于精确度的相似性度量方法.它根据分析待评估数据和参考数据中 N 元组(N-gram)共同出现的程度,来衡量机器标注数据与人工标注数据的相似性（即机器标注数据的质量）,共同出现的程度越高就代表文本标注的质量越好.(2) ChinYew 提出的 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[78]算法是一种基于召回率的相似性度量方法,与 BLEU 类似.它主要考察待评估标注的充分性和忠实性,并通过计算 N 元组在参考数据和待评估数据的共现率来评估文本标注的质量.ROUGE 算法一共有四种改进算法[79],即 ROUGE-N[80],ROUGE-L[81],ROUGE-W[82] 和 ROUGE-S[83].(3) Alon Lavie 等人提出的 METEOR[84]算法是一种基于精确度和召回率的相似性度量方法,相比于单纯基于精度的 BLEU 算法,其结果和人工标注结果的相关性较高.METEOR 算法解决了 BLEU 算法中的固有缺陷,即同义词匹配问题,以及 ROUGE 算法无法评估文本数据流畅度的缺陷.(4) Vedantam 等人提出的 CIDEr(Consensus-based Image Description Evaluation) 算法[85]是一种用于评估图像描述（Image Caption 或 Image Description）的算法,图像描述是指用一段描述性的文字说明图像中物体之间的关系.CIDEr 算法将每个句子都看作“文档”,将其表示成 TF-IDF (Term Frequency-Inverse Document Frequency)向量的形式,通过对每个 n 元组进行(TF-IDF)权重计算,比较模型生成的描述与人工描述之间的余弦相似度.如果余弦相似度越高则代表图像描述的质量越好.2017 年,研究者提出了改进的 CIDEr-D[86]算法,它通过增加截断（Clipping）和基于长度的高斯惩罚来解决 CIDEr 算法中对一个句子经过人工判断得分很低,但在自动计算标准中得分很高的情况.(5) Anderson 等人提出的 SPICE(Semantic Propositional Image Caption Evaluation)算法[87]也用于评估图像描述质量,它使用基于图的语义表示来编码标注数据中的对象、属性和关系.将待评价 Caption 和人工标注 Caption 用概率上下文无关文法依赖解析树[88]解析成语法依赖树,然后用基于规则的方法把依赖解析树映射成场景图,最后用 F-score 来评估两个场景图的相似性.如果分值越高则表示图像的描述质量越好.(6) Gianluca Demartini 等人提出的 ZenCrowd 算法[89,90]使用二元组{good,bad}来建模众包标注者的可靠性(Reliability),算法通过只使用一个参数就解决了数据稀疏性造成的变量估计偏差过大问题[91].ZenCrowd 算法计算每个众包标注者的可靠度以及使用可靠度来更新每个样本属于特

定类别的概率,标注者的可靠度越高则标注质量就越好.

5.3 语音标注质量评估算法

目前,语音标注质量评估算法主要有词错误率(Word Error Rate,WER)算法[92]和句子错误率(Sentence Error Rate,SER)算法[93].词错误率[94]表示为了让识别出来的词序列和标准的词序列之间保持一致,而需要进行替换、删除或者插入的某些词.WER 的计算公式如下:

$$WER = (S+D+I)/N = (S+D+I)/(S+D+H) \tag{2}$$

其中,S 为替换的字数,D 为删除的字数,I 为插入的字数,H 为正确的字数,N 为(替换 + 删除 + 正确)的字数.WER 的值越低表示标注效果越好,反之则表示标注效果越差.中文标注一般用字符错误率[95](Character Error Rate,CER)来表示 WER.目前,主要用字符串编辑距离(Levenshtein Distance)来计算词错误率.

SER 算法被用来识别句子中是否出现词识别错误,它的计算公式如下:

$$SER = SEN/STN \tag{3}$$

SEN(Error Number of Sentence,SEN)指句子识别错误的个数,也就是说如果句子中出现一个词识别错误,那么这个句子被认为识别错误,STN(Total Number of Sentence,STN)指句子总数.SER 的值越高就代表语音标注的质量越差;反之,则表示语音标注的质量较好.最后,表 4 对比了上述各个标注质量评估算法的优缺点.

Table 4 Comparison of evaluation algorithms for data annotation quality

表 4 各数据标注质量评估算法对比

分类	算法名称	优点	缺点
图像标注质量评估算法	MV 算法	简单易用,常用做其它众包质量控制算法的基准算法	没有考虑到每个标注任务、标注者的不同可靠性
	EM 算法	在一定意义下可以收敛到局部最大化	数据缺失比例较大时,收敛速度比较缓慢
	RY 算法	将分类器与 Ground-truth 结合起来进行学习	需要对标注专家的特异性和敏感性强加先验
文本标注质量评估算法	BLEU 算法	方便、快速、结果有参考价值	测评精度易受常用词干扰
	ROUGE 算法	参考标注越多,待评估数据的相关性就越高	无法评价标注数据的流畅度
	METEOR 算法	评估时考虑了同义词匹配,提高了评估的准确率	长度惩罚,被评估的数据量小的时候测量精度较高
	CIDEr 算法	从文本标注质量评估的相关性上升到质量评估的相似性进	对所有匹配上的词都同等对待会导致部分词的重要性被削弱
	SPICE 算法	从图的语义层面对图像标注进行评估	图的语义解析方面还有待进一步完善
	ZenCrowd 算法	将算法匹配和人工匹配结合,在一定程度上实现了标注质量和效率的共同提高	无法自动为定实体选择最佳数据集
语音标注质量评估算法	WER 算法	可以分数字、英文、中文等情况分别来看	数据量大的时候性能会特别差
	SER 算法	对句子的整体性评估要优于 WER 算法	句错误率较高,一般是词错误率的 2~3 倍

6 数据标注发展趋势

随着人工智能的兴起,深度学习、增强学习、机器学习等人工智能领域对数据标注的需求度越来越高,数据标注的重要性也不断凸显.但是,其在发展过程中也面临着一些挑战和问题,具体内容如下所示:

挑战 1:不同的行业应用对数据标注的任务存在一定的差异性,现有的标注任务还不够细化,无法满足行业的新技术需求.

现有的标注任务主要分为 5 大类,不过,随着人工智能技术的普及,一些行业对数据标注提出了更高的需求.例如,智能安防是数据标注的一个典型应用行业,常用的标注任务为图像标注中的人脸标注和行人标注.人脸标注可用于识别住户或来访者的身份,行人标注用来统计一定区域里的人群数量,并判断该区域是否出现过于拥挤的现象以避免出现踩踏事件.但是,随着技术进步,居民对智能安防系统提出了更高的需求,希望能从以往的

被动防御走向主动预警.为此,现有的标注任务已经不能满足这一需求,需要出现更加专业和更加细化的标注内容.

挑战 2:尽管数据标注工具能在一定程度上帮助标注员完成标注任务,但是整体的标注效率仍然较为低下.

在图像标注工作中,传统的人工标注方法是由标注员根据标注需求,并借助相关工具在图片上完成诸如分类、画框、注释和标记等工作.比如,在COCO + Stuff数据集中[96],标记一个图像大约需要 20 分钟,如果要标记整个数据集则需要约 53000 小时!数据标注需求的高速增长与标注效率的低效并存.

挑战 3:现有数据标注平台普遍采用众包模式来分配标注任务、造成标注结果的质量层次不齐,影响算法模型的准确性.

人工智能应用对数据标注的质量要求非常高,然而,数据标注质量的参差不齐成为人工智能企业最为苦恼的事情.现阶段数据标注主要依靠人力来完成,当标注员面临复杂的标注任务或者百万级的标注数据量时就会产生巨大的心理压力;再加上数据标注工作本身的重复性高,标注时间紧迫以及缺少严格的质量审核流程就会造成标注任务的合格率低、标注不完备[97]、标注不及时等问题.这些问题影响了后续分析结果的准确性,也会阻碍人工智能的发展过程[98].

挑战 4:基于众包模式的数据标注任务会造成用户数据缺乏安全性,并面临隐私泄露的风险.

一些金融机构和政府部门格外关注外包标注数据的安全性,但是,一些互联网企业为了降低标注成本会将用户私人社交内容标注工作层层转包给其他国家的合同工.据路透社报道,Facebook将部分的数据标注工作外包给了印度公司WiPro.该公司雇用了 260 多名工人,按照五个类别对用户发布的私人帖子进行标注.鉴于Facebook之前在数据安全上的表现,数据标注的外包行为引起了许多用户的担忧. 进而引发了用户对隐私信息泄露的忧虑

综上,在新环境和新技术下,数据标注的研究方向在于:(1)针对特定的行业需求,研究如何细化本行业的标注任务;(2)开发人工标注+机器辅助标注并存的半自动化标注工具,同时,逐步提高机器标注的占比并减少人工标注的比例;(3)研究提高数据标注质量的技术和方法;(4)研究能保证数据标注安全性和隐私性的技术和措施;下面简要介绍各研究热点所涉及的相关理论和技术.

6.1 细化数据标注任务

随着人工智能技术在一些行业的广泛应用,这些行业原有的数据标注任务已经不再满足业务需求.以智能安防为例,为了促进智能安防系统从传统的被动防御走向智能化的主动预警,一些新的数据标注任务也应运而生.例如,当一个神情紧张或者头戴面罩的小偷手握一根棍子准备翻越小区外墙企图实施盗窃行为时,安防系统应该马上启动报警系统,并及时向安防人员发出警告,以保障住户的财产安全.实现异常情况预警的新标注任务,包括表情标注、危险品标注和行为标注,利用这些数据标注就能帮助安防系统识别紧张的表情、违法的面罩和违规的翻越行为以及可能的凶器——棍子.从技术角度来看,新标注任务为异常行为的识别与建模提供了高质量的训练数据,也有利于提高模型训练的准确性.因此,针对特定的行业需求细化标注任务将是今后数据标注的一个发展趋势.

6.2 半自动化数据标注工具的研发

随着 AI 技术的发展,数据标注工具需要从只支持人工标注逐渐转化为人工标注+AI 辅助标注的方法.其基本思路为:基于以往的标注,可以通过 AI 模型对数据进行预处理,然后由标注人员在此基础上做一些校正.以图像标注为例,标注工具首先通过预训练的语义分割模型来处理图像,并生成多个图像片段、分类标签及其置信度分数.置信度分数最高的片段用于对标签的初始化,呈现给标注者.标注者可以从机器生成的多个候选标签中为当前片段选择合适的标签,或者对机器未覆盖到的对象添加分割段.AI 辅助标注技术的应用能够大大降低人力成本并使标注速度大幅提升.目前,已经有一些数据标注公司开发了相应的半自动化工具,但是从标注比例来看,机器标注占 30%左右,而人工标注占比达到 70%左右.因此,数据标注工具的发展趋势是开发以人工标注为主机器标注为辅的半自动化标注工具,同时,减少人工标注的比例并逐步提高机器标注的占比.

6.3 数据标注质量的改善

为了改善数据标注的质量,可以从以下三个层面开展相关研究.

方法一,现有的众包工作大多集中在标签推理和激励机制的设计上[99,100],今后可以考虑利用自适应群体教学（即通过监督人群以教学的形式进行标注）来提高标注质量,或者利用隐藏在“脏数据”[103,104]中的有用信息以降低标注样品（构建机器学习算法模型时用到的人工标注数据）的比例,它主要通过在脏数据上迭代地训练分类器,并根据迭代期间的估计置信度移动聚类中心,校正或删除样本.删除样本用来去除某些无法校正的低质量标注样品,以达到在保证标注样品质量的情况下降低标注样品比例,并实现对机器标注数据质量的提高.还可以利用模式识别结合一致性对标注数据进行评估并对标注人员排序以提高标注质量.

方法二,针对被标注数据数量过大的情况,可以采取自动识别和概率统计[103]相结合的方法提高对异常数据识别的效率,这里主要是指基于SOM（Self-Organizing Map）和SVM [104]的概率分布自动识别模式.SOM具有良好的矢量量化、数据融合和快速聚类能力,SVM在样本统计学习和倾向泛化方面表现良好.因此,将它们结合为两层结构模式,可以快速地识别异常数据的概率分布.同时,通过关联数据将同一类标注对象进行整合并分类管理,以便度量和监视大型标注团队的性能和质量.这样也能有效地提高数据标注的效率和质量.

方法三,将学习人群模型[105]与交互式可视化相结合,使专家能够快速访问最不确定的实例标签并去和工作人员进行验证,以此来提高标注数据的质量.

此外,如何将人类经验与学习规则充分结合以获取符合算法需求的高质量标注数据,如何对标注人员进行规范培训,如何制定标准的审核流程和控制标注质检的成本,如何从非专家提供的大量噪声标签中推断出真正的标签等都是目前数据标注质量需要尽快解决的问题和研究的方向.

6.4 数据标注中的安全性与隐私保护

为了保证数据标注平台中数据的安全性和隐私不被泄露,可以考虑采取数据治理、数据分割、数据安全传输和区块链等技术.数据治理是指对数据采集、数据清洗、数据标注到数据交付生命周期的每个阶段进行识别、度量、监控、预警等一系列管理活动,并通过改善和提高组织的管理水平确保数据在一个可控环境下使用.数据分割是指将涉密的待标注数据拆分成多个部分,分别指派给没有关联的不同团队,并且用数据接口的方式来传输数据,避免客户的数据被直接打包并互相传送,以便尽可能提高安全性.待标注的数据在分发和交付时都会涉及到数据传输,为了解决数据传输过程中存在的被盗、暴露和复制等安全性问题,就需要设计和开发出一个安全的标注数据传输框架,该框架需要提供数据加密、数据压缩和自动数据发送等功能[106].此外,基于区块链的数据标注平台采用强加密算法以及分布式技术来保障数据的安全,而且由于实现了社区自治,标注人员直接与提供标注需求的企业对接并获得标注报酬,避免标注任务的层层转包.平台一旦建设完成,全网节点均是平台的维护成员.区块链技术的使用可以避免企业用户（上传数据的账户）恶意搜集数据,也能防止个人用户（标注人员账户）批量搜集数据.

7 总结

数据标注的准确性决定了人工智能算法的有效性,因此,数据标注不仅需要有系统的方法、技术和工具,还需要有质量保障体系.本文概述了数据标注的发展,指明了数据标注目前存在的标注效率低下、标注结果的质量层次不齐、数据标注缺乏安全性以及标注任务还不够细化等问题.此外,本文还分析了数据标注未来的研究方向：（1）半自动化数据标注工具的研发,（2）数据标注质量的改善,（3）数据标注中的安全性与隐私保护以及（4）细化数据标注任务.

人工智能的终极目标是让“人工智能自主学习,自主标记,而不依赖人类对人工智能的标注与训练”[107] ,斯坦福大学已经通过一种编程方式生成训练数据的“弱监督”范式,并开发了基于弱监督编程范式 Snorkel 的开源框架:编程训练数据 (programming training data)[108].同时,编程训练数据应用于多任务学习 (MTL)场景,解决了为一个或多个相关任务提供噪声标签的问题[109].亚利桑那州立大学的研究人员提出了一种基于机器教学的自

适应交互型众包教学框架(JEDI)[110],它能够保证样本标注数据训练的有效性和多样性以及标注的质量.未来,我们期待人工智能实现真正的“智能”,能够反向作用于数据标注产业,使得人工标注逐渐转变为半自动化标注,进而向自动化标注迈进.

References:

[1] Xuan Z. Hidden "foxconn" labor-intensive industry in artificial intelligence industry. Internet weekly, 2018, 675(21):28-29. (in Chinese).

[2] Yoshua Bengio. Learning Deep Architectures for AI. Foundations and Amp; Trends in Machine Learning, 2009, 2(1):1-127.

[3] Corea F. How AI Is Changing the Insurance Landscape.2019, 10. 1007/978-3-319-77252-3(Chapter 2): 5-10.

[4] Alonso O. Challenges with Label Quality for Supervised Learning. Journal of Data and Information Quality, 2015, 6(1):1-3.

[5] Brendel W, Rauber J, Bethge M, et al. Decision-Based Adversarial Attacks: Reliable Attacks against Black-Box Machine Learning Models. International conference on learning representations, 2018.

[6] IMAGENET. http://image-net.org/about-stats

[7] Kornblith S, Shlens J, Le Q V, et al. Do better ImageNet models transfer better? Computer vision and pattern recognition, 2019: 2661-2671.

[8] Zhu J, Kaplan R, Johnson J, et al. HiDDeN: Hiding Data with Deep Networks. European conference on computer vision, 2018: 682-697.

[9] Wang C, Blei DM, Li F, et al. Simultaneous image classification and annotation. Computer vision and pattern recognition, 2009: 1903-1910.

[10] Bearman A, Russakovsky O, Ferrari V, et al. What’s the Point: Semantic Segmentation with Point Supervision. european conference on computer vision, 2016: 549-565.

[11] Debattista J, Auer S, Lange C, et al. Luzzu—A Methodology and Framework for Linked Data Quality Assessment. Journal of Data and Information Quality, 2016, 8(1): 4.

[12] Reitan E H, Saib S H. Computer graphics in an automatic aircraft landing system. national computer conference, 1976: 689-700.

[13] Kodali R K, Jain V, Bose S, et al. IoT based smart security and home automation system. international conference on computing communication and automation, 2016: 1286-1289.

[14] Liyakathunisa Syed, Saima Jabeen, S. Manimala, et al. Data Science Algorithms and Techniques for Smart Healthcare Using IoT and Big Data Analytics: Towards Smarter Algorithms// Smart Techniques for a Smarter Planet. 2019.

[15] Khan M , Wu X , Xu X , et al. Big Data Challenges and Opportunities in the Hype of Industry 4.0// IEEE ICC 2017. IEEE, 2017.

[16] Appen. https://appen.com/blog/how-ai-driving-innovation-ecommerce-retail/

[17] Zhang Y, Lu Y. Research on the Problems and Strategies of Rural E-Commerce in the Age of Internet + Agriculture. semantics knowledge and grid, 2018: 257-260.

[18] Christen P, Gayler R W, Tran K, et al. Automatic Discovery of Abnormal Values in Large Textual Databases. Journal of Data and Information Quality, 2016, 7(1): 7

[19] Sivarajah U, Kamal M M, Irani Z, et al. Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 2017: 263-286.

[20] Guo XM, Ma LL, Su K, et al. Research on Automatic Evaluation Method of Metadata Quality of Data Repositories Based on Sematic Annotation . Computer Applications and Software, 2018, 35(06):29-33, 88. (in Chinese with English abstract).

[21] Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., and Roth, D. Question answering via integer programming over semi-structured knowledge. International Joint Conference on Artificial Intelligence, 2016-January, 1145-1152.

[22] Ling H, Gao J, Kar A, et al. Fast Interactive Object Annotation with Curve-GCN. 2019.

[23] Barbosa L, Carvalho B W, Zadrozny B, et al. Pooling Hybrid Representations for Web Structured Data Annotation. arXiv: Databases, 2016.

[24] Jing K .Review of evaluation of social tagging system. Jiangsu Science & Technology Information, 2018, 35(11): 8-10. (in Chinese with English abstract).

[25] Zhang L, Wang T, Liu Y, et al. A semi-structured information semantic annotation method for Web pages. Neural Computing and Applications, 2019(5):1-11.

[26] Cai L, Liang Y, Zhu YY, et al. History and Development Tendency of Data Quality. Computer Science, 2018, 45(4): 1-10. (in Chinese with English abstract).

[27] Egorow O, Lotz A, Siegert I, et al. Accelerating manual annotation of filled pauses by automatic pre-selection// International Conference on Companion Technology. 2018.

[28] Zheng G, Mukherjee S, Dong XL, et al. OpenTag: Open Attribute Value Extraction from Product Profiles. Knowledge discovery and data mining, 2018: 1049-1058.

[29] Barthelmess P, Kaiser EC, Huang X, et al. Collaborative multimodal photo annotation over digital paper. International conference on multimodal interfaces, 2006: 4-11.

[30] Zhang B, Hao J, Ma G, et al. Automatic image annotation based on semi-paired probabilistic canonical correlation analysis. Journal of Software, 2017, 28(2) : 292-309. (in Chinese with English abstract).

[31] Vielhauer C, Schott M, Kratzer C, et al. Nested Object Watermarking: Transparency and Capacity Evaluation. Electronic imaging, 2008, 6819:18.

[32] Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2018, 34(8): 1381-1388.

[33] Pearson J, Robinson S, Jones M, et al. PaperChains: Dynamic Sketch+Voice Annotations. conference on computer supported cooperative work, 2015: 383-392

[34] Ceolin D, Groth P T, Maccatrozzo V, et al. Combining User Reputation and Provenance Analysis for Trust Assessment. Journal of Data and Information Quality, 2016, 7(1): 1-28.

[35] J. Levinson, J. Askeland, J. Becker, et al. Towards fully autonomous driving: Systems and algorithms, 2011 IEEE Intelligent Vehicles Symposium, 2011, 163-168.

[36] Hillier L W, Graves T, Fulton R S, et al. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature, 2005, 434(7034): 724-731.

[37] Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 2017, 47(1): 1-66.

[38] Liu C, Su T, Yu L, et al. Self-Correction Method for Automatic Data Annotation. Asian conference on pattern recognition, 2017: 911-916.

[39] Uijlings J R, Konyushkova K, Lampert C H, et al. Learning Intelligent Dialogs for Bounding Box Annotation. Computer vision and pattern recognition, 2018: 9175-9184.

[40] Wang C. Image annotation refinement using random walk with restarts. Acm Multimedia, 2016:647-650.

[41] Parmar B R, Jarrett T R, Burgon N S, et al. Comparison of left atrial area marked ablated in electroanatomical maps with scar in MRI. Journal of Cardiovascular Electrophysiology, 2014, 25(5): 457-463.

[42] D R Perrott, K Marlborough. Minimum audible movement angle: marking the end points of the path traveled by a moving sound source. The Journal of the Acoustical Society of America, 1989, 85 (4): 1773-1775.

[43] Best-Rowden L, Jain A K. Longitudinal Study of Automatic Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017(40): 148 – 162.

[44] Müller R H, Clegg D L. Automatic Paper Chromatography. Annals of the New York Academy of Sciences, 1949, 53(5):1108-1118.

[45] Sikos L F. RDF-powered semantic video annotation tools with concept mapping to Linked Data for next-generation video indexing: a comprehensive review. Multimedia Tools and Applications, 2017, 76(12): 14437-14460.

[46] Willis C G, Law E, Williams A C, et al. CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. New Phytologist, 2017, 215(1): 479-488.

[47] Julie McDonough Dolmaya. Analyzing the Crowdsourcing Model and Its Impact on Public Perceptions of Translation. Translator, 2012, 18(2): 167-191.

[48] Aktas A, Alexa C, Andreev V, et al. Measurement of inclusive jet production in deep-inelastic scattering at high and determination of the strong coupling. Physics Letters B, 2007: 134-144.

[49] Chen K, Chang C, Wu C, et al. Quadrant of euphoria: a crowdsourcing platform for QoE assessment. IEEE Network, 2010, 24(2): 28-35.

[50] Thomas Kohler. Crowdsourcing-Based Business Models: California Management Review, 2015, 57(4): 63-84.

[51] Chalam K V, Jain P, Shah V A, et al. Evaluation of web-based annotation of ophthalmic images for multicentric clinical trials. Indian Journal of Ophthalmology, 2006, 54(2).

[52] Tang J, Li H, Qi G, et al. Image Annotation by Graph-Based Inference with Integrated Multiple/Single Instance Representations. IEEE Transactions on Multimedia, 2010, 12(2): 131-141.

[53] Zhou H, Gao B, Wu J, et al. Adaptive Feeding: Achieving Fast and Accurate Detections by Adaptively Combining Object Detectors. arXiv: Computer Vision and Pattern Recognition, 2017(1): 3523-3533.

[54] Jongejan B. Automatic annotation of head velocity and acceleration in Anvil. Language Resources and Evaluation, 2012: 201-208.

[55] Tulasi R L, Rao M S, Ankita K, et al. Ontology-Based Automatic Annotation: An Approach for Efficient Retrieval of Semantic Results of Web Documents// Proceedings of the First International Conference on Computational Intelligence and Informatics. 2017.

[56] Xie C, Mao X, Huang J, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research, 2011: 316-322.

[57] Liu P, Zhang Y. Data annotation engineering. Tsinghua University Press, 2019. (in Chinese).

[58] Berriel R F, Rossi F S, De Souza A F, et al. Automatic large-scale data acquisition via crowdsourcing for crosswalk classification: A deep learning approach. Computers and Graphics, 2017: 32-42.

[59] Boselli R, Cesarini M, Mercorio F, et al. An AI Planning System for Data Cleaning. European conference on machine learning, 2017: 349-353.

[60] Li R, Lin Z, Lin HL, et al. Text Emotion Analysis: A Survey. Journal of Computer Research and Development, 2018, 55(1):30-52. (in Chinese with English abstract).

[61] Lei LY. Research on Fine-grained Sentiment Analysis Base on Chinese. HengYang: University of South China, 2014. (in Chinese with English abstract).

[62] Cai L, Pan J, Wei BL, et al. Visualization Analysis for Spatio-temporal Pattern of Hotspots and Sentiment Change Towards Microblog Check-in Data. Miniature microcomputer system, 2018(9):1889-1894. (in Chinese with English abstract).

[63] Cao W. Research of the Algorithm of Region-value Annotation in Crowdsouring, Nanjing: Nanjing University of Finance and Economics. 2017. (in Chinese with English abstract).

[64] Gennari R, Tonelli S, Vittorini P, et al. Challenges in Quality of Temporal Data — Starting with Gold Standards. Journal of Data and Information Quality, 2015, 6(2): 9.

[65] Xu TZ, Xu ZY. Combination Method of Fisher Theory and the Majority of Voting for Data Fusion. Science and technology Information. 2009, 27: 52-53. (in Chinese with English abstract).

[66] Wang Y, Rao Y, Zhan X, et al. Sentiment and emotion classification over noisy labels. Knowledge Based Systems, 2016: 207-216.

[67] Snow R, O'Connor B, Jurafsky D, et al. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, 2008: 254-263.

[68] Sorokin A, Forsyth D. Utility data annotation with Amazon Mechanical Turk. In: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop on Internet Vision, Anchorage, 2008: 1-8.

[69] Rahul Gupta. Modeling Multiple Time Series Annotations as Noisy Distortions of the Ground Truth: An Expectation- Maximization Approach. IEEE Transactions on Affective Computing, 2018, 9 (1):76-89.

[70] Zeng J, Liu Z, Cao X, et al. Fast Online EM for Big Topic Modeling. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 675-688.

[71] Wang W, Zhou Z. Crowdsourcing label quality: a theoretical analysis. Science in China Series F: Information Sciences, 2015, 58(11): 1-12.

[72] A. P. Dawid. Skene A M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Journal of the Royal Statistical Society, 1979, 28(1):20-28.

[73] Raykar V C, Yu S, Zhao L H, et al. Supervised learning from multiple experts: whom to trust when everyone lies abit. In: Proceedings of the 26th International Conference on Machine Learning, Quebec, 2009: 889-896.

[74] Raykar V C, Yu S, Zhao L H, et al. Learning From Crowds. Journal of Machine Learning Research, 2010, 11(2):1297-1322.

[75] Yu H, Chen Y. Clustering Ensemble Method Using Three-way Decisions Based on Spark. J. Zhengzhou University (Nat. Sci. ED.), 2018, 50(01):23-29. (in Chinese with English abstract).

[76] Vogel T, Heise A, Draisbach U, et al. Reach for gold: An annealing standard to evaluate duplicate detection results. Journal of Data and Information Quality, 2014, 5(1): 5.

[77] Papineni K, Roukos S, Ward T, et al. Bleu: a Method for Automatic Evaluation of Machine Translation. Meeting of the association for computational linguistics, 2002: 311-318.

[78] Lavie A, Agarwal A. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. workshop on statistical machine translation, 2007: 228-231.

[79] Jiang YH, Ding L, LI JE, et al. Abstractive summarization model considering hybrid lexical features. Journal of Hebei University of Science and Technology, 2019, 40(2):152-158. (in Chinese with English abstract).

[80] Jin F, Huang M, Lu Z, et al. Towards Automatic Generation of Gene Summary. North American chapter of the association for computational linguistics, 2009: 97-105.

[81] Plaza L. A semantic graph-based approach to biomedical summarization. Artificial Intelligence in Medicine, 2011, 53(1): 1-14.

[82] Campr M, Ježek K. Comparing Semantic Models for Evaluating Automatic Document Summarization. Text speech and dialogue, 2015: 252-260.

[83] Kang SZ, Hong MA, Huang R Y. An Opinion and MRW Based Sentiment Summarization Framework. Acta Electronica Sinica, 2017, 45(12):3005-3011.

[84] Lin C. ROUGE: A Package for Automatic Evaluation of Summaries. Meeting of the association for computational linguistics, 2004: 74-81.

[85] Vedantam R, Zitnick C L, Parikh D, et al. CIDEr: Consensus-based image description evaluation. Computer vision and pattern recognition, 2015: 4566-4575.

[86] Chen T, Liao Y, Chuang C, et al. Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner. arXiv: Computer Vision and Pattern Recognition, 2017 : 521-530.

[87] Anderson P, Fernando B, Johnson M, et al. SPICE: Semantic Propositional Image Caption Evaluation. European conference on computer vision, 2016: 382-398.

[88] Cui Y, Yang G, Veit A, et al. Learning to Evaluate Image Captioning. Computer vision and pattern recognition, 2018: 5804-5812.

[89] Demartini G, Difallah D E, Cudré Mauroux, Philippe. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking// International Conference on World Wide Web. ACM, 2012: 469-478.

[90] Zhang J, Sheng V S, Wu J, et al. Multi-Class Ground Truth Inference in Crowdsourcing with Clustering. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(4): 1080-1085.

[91] Ruckhaus E, Vidal M, Castillo S, et al. Analyzing Linked Data Quality with LiQuate. European semantic web conference, 2014: 488-493.

[92] Ruiz N, Federico M. Phonetically-oriented word error alignment for speech recognition error analysis in speech translation.// Automatic Speech Recognition and Understanding. 2016: 296-302.

[93] Augello A, Cuzzocrea A, Pilato G, et al. An Innovative Similarity Measure for Sentence Plagiarism Detection. International conference on computational science and its applications, 2016: 552-566.

[94] Spiccia C, Augello A, Pilato G, et al. Semantic Word Error Rate for Sentence Similarity. IEEE international conference semantic computing, 2016: 266-269.

[95] Escudero J P, Novoa J, Mahu R, et al. An improved DNN-based spectral feature mapping that removes noise and reverberation for robust automatic speech recognition. arXiv: Audio and Speech Processing, 2018.

[96] Andriluka M, Uijlings J R, Ferrari V, et al. Fluid Annotation: A Human-Machine Collaboration Interface for Full Image Annotation. ACM multimedia, 2018: 1957-1966.

[97] Yang B, Kaul M, Jensen C S. Using Incomplete Information for Complete Weight Annotation of Road Networks. 2014: 1267-1279.

[98] Marcheggiani D, Sebastiani F. On the Effects of Low-Quality Training Data on Information Extraction from Clinical Reports. Journal of Data and Information Quality, 2017, 9(1): 1-25.

[99] Maccartney B, Galley M, Manning C D, et al. A Phrase-Based Alignment Model for Natural Language Inference. Empirical methods in natural language processing, 2008: 802-811.

[100] Gururangan S, Swayamdipta S, Levy O, et al. Annotation Artifacts in Natural Language Inference Data. arXiv: Computation and Language, 2018: 107-112.

[101] Wang Z, Xu Y, Suo B, et al. A Provenance Storage Method Based on Parallel Database. International conference on information science and control engineering, 2015: 63-66.

[102] Liu C, Su T, Yu L, et al. Self-Correction Method for Automatic Data Annotation. Asian conference on pattern recognition, 2017: 911-916.

[103] Skala W, Wohlschlager T, Senn S, et al. MoFi: A Software Tool for Annotating Glycoprotein Mass Spectra by Integrating Hybrid Data from the Intact Protein and Glycopeptide Level. Analytical Chemistry, 2018, 90(9): 5728-5736.

[104] Zhu J, Zhang H, Guo J, et al. Data distributions automatic identification based on SOM and support vector machines. International conference on machine learning and cybernetics, 2002: 340-344.

[105] Wang J. A review of China's statistical data quality research //Fortune Today Forum. 2016: 393-394. (in Chinese with English abstract).

[106] Yang Y, He H, Wang D, et al. A Framework to Data Delivery Security for Big Data Annotation Delivery System. Mobile adhoc and sensor systems, 2018: 532-536.

[107] Verhulst S G. Where and when AI and CI meet: exploring the intersection of artificial and collective intelligence towards the goal of innovating how we govern. Ai & Society, 2018(5): 1-5.

[108] Alex Ratner, Paroma Varma, Braden Hancock, et al. Weak Supervision: A New Programming Paradigm for Machine Learning //https://ai.stanford.edu/blog/weak-supervision/. The Stanford AI Lab Blog, 2019.

[109] Shetty R, Fritz M, Schiele B. Adversarial Scene Editing: Automatic Object Removal from Weak Supervision. 2018.

[110] Zhou Y, Nelakurthi A R, He J. Unlearn What You Have Learned: Adaptive Crowd Teaching with Exponentially Decayed Memory Learners. 2018.

附中文参考文献:

[1] 轩中. 人工智能行业中隐藏的“富士康”式劳动密集型产业. 互联网周刊, 2018, 675(21):28-29.

[20] 郭晓明,马良荔,苏凯等.基于语义标注的数据资源库元数据质量自动评估方法研究.计算机应用与软件,2018,35(06): 29-33,88.

[24] 敬凯. 社会化标注系统评价研究述评. 江苏科技信息, 2018, 35(11):8-10.

[26] 蔡莉, 梁宇, 朱扬勇等. 数据质量的历史沿革和发展趋势.计算机科学, 2018, 45(4):1-10.

[30] 张博, 郝杰, 马刚等. 基于弱匹配概率典型相关性分析的图像自动标注. 软件学报, 2017, 28(2): 292-309.

[57] 刘鹏, 张燕. 数据标注工程. 清华大学出版社, 2019.

[60] 李然, 林政, 林海伦等. 文本情绪分析综述. 计算机研究与发展, 2018, 55(1):30-52.

[61] 雷龙艳. 中文微博细粒度情绪识别研究. 衡阳: 南华大学, 2014.

[62] 蔡莉,潘俊,魏宝乐等. 签到数据的热点区域时空模式与情感变化的可视化分析. 小型微型计算机系统, 2018, 39(9):1889-1894.

[63] 曹伟. 众包域值标注算法研究, 南京:南京财经大学, 2017.

[65] 徐太征,徐中宇. Fisher 理论和多数投票法相结合的数据融合算法. 科技信息. 2009, 27: 52-53.

[75] 于洪, 陈云. 基于 Spark 的三支聚类集成方法. 郑州大学学报(理学版), 2018, 50(1): 20-26.

[79] 江跃华,丁磊,李娇娥等. 融合词汇特征的生成式摘要模型. 河北科技大学学报,2019,40(2):152-158.

[105] 王晶. 中国统计数据质量研究综述.今日财富论坛. 2016: 393-394.