浅谈数据标注产业的现实发展意义

郭 凯

（山西省晋中市左权县职业技术中学校 山西 晋中 032600）

[摘 要] 随着社会主义现代化建设进程的不断加快，计算机技术的不断发展和广泛应用在一定程度上促进了数据的“井喷式”增长，进而为大数据时代来临奠定了良好基础。而服务于大数据分析处理的数据标注产业作为基础性工程也定将长期发挥重要作用，对于产业进步和经济发展具有重要现实意义。本文立足大数据产业时代，对数据标注行业的现状和未来发展进行介绍分析，并浅谈其现实发展意义。

[关键词] 大数据；数据标注；感知

数字经济是全球新一轮科技和产业革命最典型的标志，其中最关键的动力来自人工智能等前沿技术的创新突破。近年来，得益于人工智能的兴起，数据标注产业应运而生，它主要是根据人工智能企业的要求，对图像、声音、文字等进行不同方式的标注，从而为人工智能企业提供大量的数据供机器训练和学习。发展数据标注产业，对我省绿色健康发展转型，有着重大现实意义。

一、数据标注——机器感知世界的起点

数据标注是人工智能产业的基础，是机器感知现实世界的起点。从某种程度上来说，没有经过标注的数据就是无用数据。机器识别事物主要通过物体的一些特征。被识别的物体还需要通过数据标注才能让机器知道这个物体是什么。

在机器的世界里，图像与语音、视频等一样，是数据的一个种类。近年来，随着数码产品以及存储技术的迅速普及和发展，人们每天都可通过相机、可视电话、监控及医疗设备等制造大量图像。因此，现阶段图像已然成为标注产业发展的重点对象。

如果素材是一张人物图像，那么需要标注的信息往往是性别、面部朝向、人种、有无帽子眼镜等，也可以人为地将人物和背景的区域划分开来。将成千上万张经过标注的图片组成的数据集“投喂”给机器，它才能在一张全新的图像中分辨出人物在哪个区域、具有怎样的外貌特征。对于人来说“小儿科”的思考历程，机器却需要大量的标记数据集进行训练。

二、机器学习——缓解人工标注的压力

提到人工智能产业，人们往往联想到繁华的城市和干练的IT精英，但实际上，支撑起人工智能的数据标注产业，却是一个劳动密集型产业。百度搜索“数据标注”，会出现很多图片语音视频数据采集、标注公司。随机选择一个此类词条点进去，往往会看到“万人数据标注团队”等类似宣传语。可见人工标注是目前数据标注的主要方式。

谷歌推出的流体标注模型主要利用人工智能学习的基础，对图像数据进行自动标注，对于标注不准确或者出现偏差的地方可以通过人工调整，从而提高标注效率。即便该模型可借助机器学习提升标注速度，但最初还需进行人为地数据标注，为其提供初始训练数据集。事实也正是如此，为了标注图片，谷歌预先以约一千张具有分类标签和信任分数的图片训练了语意分割模型。

但该模型尚不完美，物体边界标记问题、界面操作速度以及类别扩展等仍需进一步研究或完善。

三、人才培养——行业发展的必经之路

大数据时代的来临，数据标注行业的快速发展，从某方面来讲，对专业人才也提出了更高要求，而计算机人才作为计算机各项作业的基础，加强培训和训练非常重要。从目前来看，企业在发展过程中，受传统发展理念根深蒂固的影响，对于人才培养的重视度较低，导致在后期发展过程中，各项作业开展都缺乏专业性人才作为指导，从而给整体经济效益和社会效益带来极为不利的影响。因此，为从根本上有效解决上述问题，在企业发展过程中，提高对人才培养的重视程度，为员工打造与自身能力相适应的培训是提升员工专业能力和综合素质的重要基础和根本前提。

对于数据标注行业而言，信息采集、筛选和加工是计算机信息处理技术的基本职能，因此，在大数据时代背景下，为从根本上推进企业进一步发展，提高信息采集、筛选速度，要不断提高计算机信息处理技术水平，确保计算机各项作业顺利开展。数据采集工作就是实时、动态化控制目标数据源，因此，在采集数据时，采集工作人员不仅需标明信息的目标源，为后期跟踪作业奠定良好基础，还要在信息录入的过程中，按照一定顺序对上述数据进行加工、处理，从而为数据利用和输送创造良好条件。

四、政策支持——为产业发展插上翅膀

《山西省人民政府关于加快我省数据标注产业发展的实施意见》中特别指出，人工智能是新一轮科技革命和产业变革的重要驱动力量，是我省重点培育的高成长性新兴产业。数据标注作为人工智能的基础环节，对于推动产业集聚发展，培育人工智能产业，促进经济结构转型具有重要意义。

《意见》同时制订了发展目标，即到2022年，引进培育100家以上数据标注企业，就业人员规模超过1万人，初步形成集数据采集、数据清洗、数据标注、数据交易、数据应用为一体的基础数据服务产业体系，初步建成涵盖无人驾驶、工业质检、医疗服务等领域的基础数据开放平台，人工智能创新应用生态初步显现，数据标注产业年产值达20亿元。到2025年，基础数据服务体系基本完善，人工智能基础数据开放平台影响力大幅提升，山西成为全国领先的基础数据产业聚集地，数据标注产业年产值达到50亿元，基础数据服务产值达到150亿元，带动人工智能相关产值达到500亿元。

随着以百度山西数据标注产业项目为代表的数据标注基地的建设，数字经济为我省实现新旧动能转换、打造绿色发展升级版带来了前所未有的机遇。

五、现实意义——拉动经济并增加就业

1.有利于挖掘新的经济增长点。据百度研究院估算，数据标注产业规模有望在2020年超过500亿元规模，阿里、京东等巨头都已推出数据标注外包业务。百度众测平台在2017年就发布了5000万元的数据标注任务，今年预计将达3亿元。未来，数据标注产业会有爆发式的增长，市场规模将进一步扩大。在我省与百度深化战略合作的大背景下，我们应抢抓机遇，提供更专业化、集约化的数据外包服务，进一步丰富数字经济产业链，挖掘新的经济增长点。

2.有利于夯实数字产业基础。我省正大力推进政府的数字化转型，以政府的数字化带动一批产业数字化，并将云计算、大数据、人工智能等列为发展重点，而这些重点产业都需要经标注的海量数据供机器训练和学习，并随着人工智能等产业的发展，新的标注需求还将不断出现。在经济数字化转型已步入人工智能驱动的转型阶段大背景下，数据标注已成为人工智能等产业链的重要基础。

3.有利于扩大就业。这一行业对年龄、学历没有太高的要求，经过专业的培训一般就可上岗，而且就业形式灵活，可全职，也可兼职。发展数据标注产业，不仅可为一部分人群提供就业机会，还能吸引外出的打工人员回流、周边的劳动力来我省工作，进一步丰富我省的人力资源结构。

4.有利于推动精准扶贫工作。对于一些贫困家庭或者是因残疾致贫的家庭，数据标注可为他们提供一份体面的收入。如聋哑人因为无法避险、难以沟通，就业较为困难，但是这类群体专注、对视觉信号敏锐，数据标注行业就比较适合他们。数据标注为新时代下推进精准扶贫工作提供了有力抓手。

数据标注是人工智能等产业不可或缺的一环，逐步呈现专业化、外包化、集约化趋势。要抢抓机遇，形成产业优势，为推动我省经济高质量发展贡献数据的力量！

参考文献

[1]徐诚瑞.发展数据标注产业的现实意义探讨[N].衢州日报，2018-09-29（007）.

[2]发展数据产业：打造山西智能时代核心优势[N].山西政协报，2018-12-14（002）.