

从人工智能角度谈语言类学科发展和人才培养*

宗成庆

(中国科学院自动化研究所 北京 100190)

摘要:近年来,以大语言模型为代表的人工智能(AI)技术快速发展,对语言类学科甚至整个人文类学科的招生、教学和科研工作带来了极大冲击,外语专业首当其冲,一时间外语学科和专业变得“岌岌可危”。在此背景下语言类学科和专业的教学和研究面临怎样的机遇和挑战?语言类学科如何走出困境,重塑自信,成为备受关注的热点话题。本文首先从人工智能技术发展的角度,以外国语言文学专业为例,分析语言学知识和语言学家在人工智能技术发展过程中的作用变化,然后指出目前外语教学和人才培养中存在的七个问题,最后给出外语学科和专业发展的六点建议。

关键词:人工智能;外语教学;语言研究;外语人才培养;人文类专业发展

中图分类号:H0 文献标识码:A 文章编号:1004-6038(2024)05-0001-11

DOI:10.13458/j.cnki.flatt.005064

1. 人工智能技术的发展

人工智能(Artificial Intelligence, 简称AI)概念于1956年被提出,近70年来经历了三次大规模的潮起潮落。其第一次兴起是在AI概念被提出之后大约10年里,这一阶段中,以专家系统为代表的人工智能技术发展迅速,在辅助决策和推理等方面发挥了较好作用。这一时期采用的主要是符号主义方法,或称“理性主义方法”,其基本思路是利用符号和规则建立知识库,基于规则和逻辑推理实现人类智能行为的模拟。由于这种方法过于依赖人工编写的规则和知识库,对于复杂问题和大规模数据的处理能力有限,推理效率低,知识获取过程复杂,速度慢,因此,20世纪60年代中后期,AI技术发展进入了第一个低迷时期。

20世纪70年代中后期大规模和超大规模集成电路的出现,使得计算机硬件性能大幅度提高,无论是存储容量,还是运算速度都实现了飞跃性提升。同时互联网出现,也使得各种应用需求急速增加。为了改进符号主义方法的缺陷,提高AI技术的性能以满足不断增长的应用需求,人们开始探索数据驱动的方法,或称“经验主义方法”。自20世纪70年代末、80年代初开始,经验主义方法开始崛起,从而推动了AI技术复苏,至1990年前后进入第二次高潮。这种方法的主要思路是首先收集大规模真实的样本(对于自然语言处理而言,“数据”就是语料),通过人工或人工辅助的手段对样本进行标注,提取数据特征,然后在大规模标注数据上进行模型训练。n元文法模型(N-Gram Model)、隐马尔可夫模型(Hidden Markov Model, 简称HMM)、支持向量机(Support Vector Machine, 简称SVM)和最大熵模型(Maximum Entropy Model, 简称MEM)等,都是这一时期的代表性模型。这些技术和方法在语音识别、

* 基金项目:本文系国家自然科学基金重点项目“大规模多语种多模态神经机器翻译关键技术研究”(项目编号:62336008)的阶段性成果。

作者简介:宗成庆,教授,博士,研究方向:自然语言处理,机器翻译,语言认知计算。作者邮箱:cqzong@nlpr.ia.ac.cn

图像识别和机器翻译等任务上获得了突破性进展,带动了相关技术的大规模产业化应用。经验主义方法的优点在于,它可以避免符号主义方法由于人工建立规则和知识库而带来的主观性因素,减轻了人工介入的工作量,缩短了系统研发周期,但是,系统的开发和推广应用成本较大(主要来自数据收集、标注、模型训练和算力),在很多应用任务上的性能表现欠佳,而人们对AI技术的期望值又过高,AI研究因此再度走向低谷。

21世纪初期,以G. E. Hinton等为代表的一批机器学习专家对神经网络模型的研究取得突破性进展,建立了深度学习方法,再次点燃了人们对AI技术的希望,从而开启了AI研究第三次崛起的序幕。深度学习方法建立在多层神经网络基础之上,被称为“连接主义方法”。这种方法的优势不仅在于其性能卓越,更在于人工参与的工作量大幅减少,不需要对大规模训练样本进行标注,模型就可以自动从数据中学习知识,人工只需要标注少部分样本用于模型微调(fine-tuning),就可以使模型性能得到快速优化。深度学习方法使图像识别的准确率得到了显著提升,语音识别的错误率大幅度下降。2011年1月IBM Watson人机对话系统在美国挑战游戏电视节目《危险边缘(Jeopardy!)》中战胜两位人类冠军,同年,苹果公司研发的iPhone 4手机的Siri语音助手允许用户直接通过语音获取信息或操控手机,手机能够自动理解人的意图,完成大部分任务。2016年12月,AlphaGo战胜围棋世界冠军李世石(Lee Sedol)。这一系列世界性重大的科技事件不断颠覆人们对AI技术的认知,同时也深刻地影响了全球化社会的发展。

2022年11月底美国OpenAI公司研发的ChatGPT公开发布,其出色的性能吸引了全世界关注的目光,由此掀起了大语言模型(Large Language Models,简称LLMs)的激烈竞赛。2023年3月14日OpenAI公司发布GPT-4.0,2024年2月15日发布Sora,5月13日发布GPT-4o(Omni),每隔几个月都有一款新的大模型产品研制成功,让整个世界目不暇接。国内从事AI技术研发的科研机构、大学和企业也不甘示弱,文心一言、通义千问、混元、百川等一大批“中国制造”的大语言模型相继被推出,AI技术由此步入大模型时代。因此,有人称“大模型时代”为AI的第四次浪潮。可以预见,在未来三到五年的时间里,AI技术对各领域、行业的颠覆性影响将成为不可逆转的事实。

2. 语言学在AI发展中的作用

1956年人工智能概念被提出时,自然语言理解(Natural Language Understanding,简称NLU)就成为该领域研究的核心问题之一,其目的是探索人类自身语言能力和语言思维活动的本质,研究模仿人类语言认知过程的自然语言处理方法和实现技术。20世纪70年代,随着互联网技术的出现和推广,人们更多地从语言技术应用的角度思考如何建立自然语言理解模型,并服务于实际应用,于是出现了自然语言处理(Natural Language Processing,简称NLP)术语(宗成庆2013)。

机器翻译(Machine Translation,简称MT)是自然语言处理中最经典的问题,它的提出比人工智能概念大约早了九年,其基本理念早在1947年就已经产生。由于机器翻译几乎涵盖了自然语言处理面临的所有问题——从词法、句法、语义到语用,从源语言的分析到目标语言的生成,无所不及,因此,它被认为是自然语言处理中,甚至是整个人工智能领域中最具挑

战性的研究课题。机器翻译方法的变迁,可以看作是整个人工智能领域技术发展历程的缩影。70多年来,机器翻译方法实现了从理性主义到经验主义,再到连接主义的过渡。图1中的曲线刻画了机器翻译技术70多年的发展史。

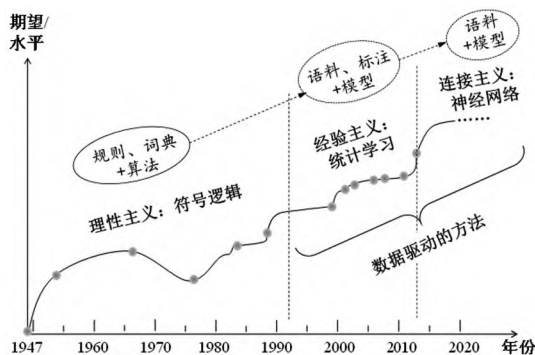


图1 机器翻译发展历程示意图

图1中两条垂直虚线将1947年至今的70多年划分成了三个基本阶段,分别对应理性主义时期、经验主义时期和连接主义时期。每个阶段上方的椭圆形里描述的是对应时期机器翻译方法的基本原理。理性主义阶段的机器翻译主流方法是基于规则的方法,其理论基础是语言学家或形式语言学家建立的结构句法理论(Tesnière 1959)、句法结构理论(Chomsky 1957)和句法翻译框架理论(Yingve 1957)等。在这一时期,语言学家是研发机器翻译系统的核心成员,他们需要精通被翻译的源语言和最终翻译产出的目标语言,利用符号逻辑手段将两种语言的词法、语法和语义等相关知识尽量准确地表示出来,并设计出源语言句子结构分析规则、两种语言之间的结构转换规则和目标语言以外生成规则。计算机专家根据语言学家建立的规则和知识库设计分析算法,最终实现两个特定语言对的机器翻译系统。在整个机器翻译系统实现和维护过程中,语言学家的作用是举足轻重的,知识库和各种规则设计的质量如何,直接影响翻译系统的译文质量和运行速度。一旦需要更换语言对,几乎所有的工作都要从头开始,而精通两种语言,且愿意为开发机器翻译系统长期服务的语言学家,是可遇不可求的。

20世纪70年代末至80年代初,基于隐马尔可夫模型实现的自动语音识别系统,以及1989年IBM公司研究建立的统计机器翻译(Statistical Machine Translation, 简称SMT)系统,可以看作是经验主义方法的标志性成果。统计机器翻译方法不再将机器翻译系统建立在语言学家构建的知识库和规则库之上,而是通过收集大规模句子级双语对照的平行语料,由机器学习(machine learning)方法自动计算出两个句子之间不同词语或者短语的对齐概率,从而推算出给定的源语言句子可能的最大概率的目标译文。为了实现这种方法,研发人员的主要任务是收集大规模高质量的双语平行语料,建立和训练机器学习模型。在这种方法刚投入实践时,基于词的翻译模型(word-based translation model)还需要熟知源文和译文两种语言的语言工程师标注双语平行句对中词语之间的对齐关系,到后来,基于短语的翻译模型(phrase-based translation model)出现之后,手工标注语料已不再是必须要做的事,尤其对于资源丰富的语言对而言,只要有足够多的双语对照语料和一个高质量的双语对照词表,模型就可以自动实现词语对齐,并计算出翻译模型的概率。20世纪90年代末至2010年的10多年里,一批词语对齐工具、翻译模型和双语平行语料等关键技术和数据相继开放之后,一大批

机器翻译系统便如雨后春笋般涌现,任何人都可以利用这些开源工具和语料构建自己的翻译系统,机器翻译技术快速走向了大规模的产业化应用。

曾有传言说,美国著名语音识别专家贾里尼克(F. Jelinek)在谈到语言学家对语音识别系统研发的作用时说过这样的话:“每当我开除一个语言学家,语音识别系统就更准了。”(冯志伟 2018)尽管这句话没有正式的文字记载,但统计机器翻译方法和语音识别技术已经在很大程度上摆脱了对语言学家的依赖,却是不争的事实。

如果说统计机器翻译系统的性能还不尽人意,译文质量远远达不到人类译员的要求,那么,2014年“序列到序列”的神经机器翻译方法的出现彻底改变了这一局面,尤其是2017年自注意力机制和“端到端”的Transformer架构被提出之后,机器翻译的译文质量有了质的提高。大量的对比实验表明,基于深度学习方法的机器翻译系统在大多数语言对上的译文质量高于翻译专业本科生,甚至不亚于硕士研究生的翻译水准。更为重要的是,不少神经机器翻译系统的开发工具已经开源,任何熟悉编程的人,只要建立一个双语对照词表,并有足够多的语料,甚至不需要太多的双语平行对齐语料,而是源语言和目标语言的单语语料,都可以实现一个神经机器翻译系统。也就是说,随着技术的发展,机器翻译系统的研发对语言学家的依赖程度进一步减少了。当然,我们必须承认,实现高质量的机器翻译系统仍然离不开外语专家的帮助,但外语专家的作用已经不再如从前一般举足轻重。

机器翻译技术的现状如此,语言分析(包括词法分析、句法分析和篇章分析等)、语料标注、自动写作、文本纠错、人机对话系统等各应用技术的现状也几乎大同小异。至此,语言类专业面临的挑战,以及语言类专业从业者感受到的压力和担忧,已成为无法回避的现实。

3. 语言类学科存在的七大问题

正如前文所述,随着AI技术的快速发展,语言学和语言学家在应用技术研发中的作用被逐步削弱,从语言学家隐约预感到的“兵临城下”到现今的“大敌当前”,仅仅过了10年左右。如果说AI技术的快速发展是导致语言类学科“岌岌可危”的外部因素,那么,从事语言类学科建设和专业教学、研究及人才培养的业内人士是否有必要审视一下学科本身的问题?笔者作为一个工科出身的门外汉,无意冒犯历史悠久、队伍庞大的语言类学科和语言学家,但从现实本身考虑,奔着有利于学科发展的目的,斗胆提出语言类学科目前存在的几方面问题。

3.1 部分教师学科理念存在偏差

这里所说的“理念”是指对学科存在和发展目标的认识。笔者认为,任何一个学科存在和发展的目的无非有两个:一个是揭示和解释某些基本规律和现象,简单来说,就是从事科学层面的研究和发现工作;另一个目的就是将学科知识应用于实际场景,解决具体问题,服务于人类社会。无论是前者,还是后者,最终的目的只有一个:学以致用。任何一项研究工作,如果没有应用价值,无论是理论的,还是实践的,都是无用的。当然,应用可以是直接的,也可以是间接的。语言类学科也应遵循这一常理。可是,部分教师对这一问题的理念存在很大的偏差,这些偏差通常表现为如下两种方式:

(1) 为研究而研究,为写论文、写专著而研究,从来不考虑应用的问题,好像应用是别人的事,可是这样“纯粹”的研究往往脱离实际,闭门造车。例如,某种语言现象被连篇累牍地分析、

解释,可是,研究者(论文作者)从来不考虑这种语言现象在现实生活中使用的概率有多大,什么样的人群、多大比例的人群在使用。如果这些基本问题和前提没有得到清楚的回答,那么这种研究的意义何在呢?当前世界已经进入全球化信息时代,互联网和移动终端上各类社交平台的“线上语言”,在很大程度上已经成为民众的通用语言,产生了太多的语言问题。如果大量真实存在的语言相关问题不能快速、有效解决,那么再“纯粹”的研究又有什么意义呢?

(2) 部分老师理所当然地认为,语言学研究的成果或者语言学家对语言现象分析总结的规律就应该被从事计算机或人工智能技术研发人员所接受和使用。这种理念或许是因为人工智能发展的早期由理性主义思维驱动的原因,但是他们却不知道计算机科学家在经历过大量的实验验证之后,已经从离散的符号逻辑思维走向了连续的实数空间表示,而语言学归纳总结的知识表示方式还停留在形式语言阶段,甚至很多语言学研究的成果根本就难以被形式化表示,而无法形式化表示的知识自然就无法被计算机技术采用。

教师理念不仅仅影响到教师本身的科研模式,更重要的是直接关系到本科教学和研究生人才培养。研究生论文从选题到研究过程和成果,脱离实际应用需求,跟不上技术发展的节奏,这就不足为奇了。

3.2 教师队伍结构单一,跨学科合作流于形式

在很多高校的语言类学科的院系里,教师队伍基本上都是本学科专业背景的,很少有其他相关专业和计算机专业背景的教师,而很多面向应用需求的研究方向需要多学科合作,如社会学、国际传播、心理学等,尤其需要与人工智能、计算机专业密切合作。但实际情况是,由于学科背景的差异,来自不同学科和不同专业的教师对同一问题、同一概念的理解存在很大差异,甚至不能在同一频道上进行对话。例如,对于“生成式语言模型”(Generative Language Model)和神经机器翻译(Neural Machine Translation, 简称NMT)概念的理解,计算机专家与语言学家可能完全不同。究其根源,笔者认为有两方面原因,一方面是在很多高校里跨学科合作流于形式。尽管学校和院系不断强调跨学科交流与合作的重要性,但在实际执行过程中的合作往往缺乏实质性内容,相当多的合作只限于开开会、听听报告、讲讲故事,而不是坐下来面向具体目标,针对具体问题,一步一步地、长期地共同研究、探讨和实施。另一方面,教师的知识结构未能跟上时代的发展和应用需求,许多老师所掌握的知识始终停留在自己上学期间所学习的内容,对于最新技术的了解主要来自道听途说,学校或学院未能给教师(尤其是青年教师)提供进修学习的机会,教师自己也缺乏主动学习的积极性,无论是由于自己因学科背景的原因而产生的“讳疾忌医”心理,还是因为繁忙的工作占用了他们的大部分时间,最终的结果就是知识结构与时代脱节。部分教师自己从来没有实践过给学生讲授的课程内容,且不说AI相关的技术,即使有些从事翻译课程教学的教师,自己也从来没有翻译过一本书。

3.3 学科划分过细,而应用需求错综复杂

近年来正在实行的高考政策为消除过早划分文理科方向所带来的负面影响发挥了重要作用。但是,在高校里学科专业却有细致的划分,研究方向更是细上加细,而且不同学科专业之间普遍存在壁垒,且不说文、理、工、医、农、林等大的学科方向之间的壁垒,即使同一类学科的不同专业之间同样存在壁垒。这一问题除了历史原因之外,与教师队伍等因素也有

密切关系。对此本文3.2节已有阐述,彭青龙(2024)也有详细论述,在此不再赘述。

与学科专业划分过细相对应的是,在实际应用中遇到的问题往往不是某个单一学科或专业能够解决的,而是错综复杂的,甚至难以分辨究竟需要哪些学科或专业知识来解决这些实际问题。例如,我们需要实时了解某个国家的民众对孔子学院的评价和反映,需要实时调整我们的政策,并且希望清楚地了解孔子学院所在大学校长的个人情况以及与其他大学校长之间的关系、政治倾向以及宗教信仰等。从自然语言处理的角度看,要实现这样一个系统,至少涉及如下技术:(1)网络数据获取,实时获取和处理与所关注的孔子学院及所在大学相关的网络内容(数据);(2)知识库(知识图谱)构建,获取孔子学院所在国家相关的知识(包括社会伦理、宗教、政治等)以及该孔子学院相关的背景信息等,建立知识库;(3)机器翻译,将获取的外文数据自动翻译成中文;(4)信息抽取,从所获取的数据中抽取与所关注的孔子学院及所在大学相关的关键信息,包括人物、事件、时间、地点、相互关系、主题等;(5)情感分析,分析相关文本的情感(观点)倾向;(6)推理分析,分析某个事件或某个观点可能引发的关注(舆情)及后果,以及建议采取的措施(如果需要);(7)信息推荐,根据该孔子学院的具体情况向教师、学生或行政管理人员推荐语言、文化、历史或政治等方面的相关知识;(8)自动摘要及报告生成,对获取的信息进行归纳总结,形成摘要,基于分析结果撰写报文,提供给用户。

从上述分析可以看出,实现该应用系统需要自然语言处理专家和具有语言类学科背景,甚至历史、社会学和文化传播等背景的专家共同完成,尤其在知识库构建、翻译、文化传播、分析结果评价等方面,具有区域国别学、数字人文和外语等相关专业背景的人才拥有充分的发挥作用的空间。当然,具有这类背景的人才需要了解人工智能专家研发系统的思维方式和具体需求,用他们能够接受的表达方式和相同概念的术语进行交流和沟通,这就要求我们从人才培养的源头考虑学科建设、课程设置和教学内容及教学方式,面向实际需要,有的放矢地开展教学和科研活动。

3.4 教材体系不够完善,教学内容未能与时俱进

教材是本科和研究生教育重要的依据,给学生讲授什么内容,采用什么样的方式讲授,对学生能力的培养具有直接影响。可是,据笔者了解,目前很多专业课程的教材并不理想,要么教材内容陈旧,几十年前编写的教材仍在使用;要么根本没有教材,全靠老师“口口相传”。对于前者,如果是基础课,笔者并不排斥使用传统的经典教材。但是,对于专业课,尤其是与实际应用相关的课程内容,如果不结合最新的技术发展情况来授课,最终变成纸上谈兵,恐怕是在所难免的。而对于后者,连基本的教材都没有,仅靠师生口口相传,是完全不可接受的,可目前不少非通用语种的授课都采用这种方式。例如波斯语、斐济语、索马里语等,从事这些语种教学工作的老师本身也不会专门学习和研究这些语言——他们被派遣到以这些语言为官方语言的国家或地区,经过几年的学习后回到国内,就开始从事这些语言的教学工作,没有教材,甚至没有一个完整的、规范的外汉双语对照词表。一旦老师调离学校,其他老师又不能及时接手,那么相关外语专业只能寿终正寝。这样不仅无法系统而长远地培养相关语种的高水平人才,就连基本的语言教学恐怕都很难做好。基于这种状况,除了几个通用语种之外,大多数非通用语言领域竟然拿不出一个基本的、高质量的、数字化的中外双语对照词表,这着实令人惊愕。

3.5 教学平台金玉其外,科研设备未能充分发挥作用

近年来,各高校都投入了大量资金对教学设备进行改造升级,教室和听音室里都安装了性能较好的电脑、大屏幕高清显示器和质量优良的音视频设备等,但拥有昂贵的硬件设备并不意味着建立了先进的教学平台。笔者认为,先进的教学平台至少应具备如下基本功能:符合专业特点,满足专业需求,便于学习,操作简单,能够为相关研究提供数据支撑。从目前情况看,大部分设备只具备某些方面的功能,因为很多设备并不是专门为教学研发的,而且有些设备或系统操作复杂,例如,被广泛用于翻译教学的计算机辅助翻译(Computer-Aided Translation, 简称 CAT)软件“塔多思”(Trados)。针对这种状况,有条件的院校完全可以采用专门为翻译专业学员培训和教学研发的技术平台,或者根据自己学校专业的需求寻求人工智能技术公司定制,建立具有专业特色的教学平台。

除了教学平台以外,很多高校语言类专业教师的科研设备往往也未能充分发挥作用。从客观上讲,很多硬件设备(主要是计算服务器或存储服务器)相对落后,除了教师自己购置的笔记本电脑外,院系普遍资金紧张,很难有足够的经费购买高性能服务器。或许由于历史的原因,有人认为从事人文类学术研究的人员不需要性能优良的计算设备,长久以来人文类学科开展的多数研究工作也的确没有基于大规模数据的实验分析和验证,这种历史状况和思维惯性,导致了人文类学科很难获得较大额度的经费支持。在综合性的或者以理工科为主的大学里,人文类学院要想获得大额的经费支持几乎是不可能的。一个教师靠着一台普通的笔记本电脑,除了上网,写课件或文稿,还能做什么呢?从教师队伍的主观上讲,是否有意愿主动学习最新的人工智能技术呢?即使提供了先进的计算设备,是否有能力用得起来呢?这也是值得反思的问题。

3.6 缺少合理有效的数据共享机制和管理平台

数据共享和管理平台是一个学科专业从事科研工作的重要资源,在学科发展中至关重要。然而令人遗憾的是,中国的很多学科领域都没有建立起合理有效的数据共享机制和管理平台。每家从事相关技术研究的单位或团队都有自己的数据,标注规范也往往各自为政,没有权威的国家资源库。以汉语分词和词性标注为例,20世纪80年代初至2010年前后的大约30年(正值经验主义方法的盛行期)是汉语自动分词方法和词性标注方法研究的鼎盛期,在这一时期很多高校、研究所和公司都做了大量的研究工作,国家也出台了《信息处理用现代汉语分词规范》(刘开瑛 2000;刘源等 1994),标注了大量语料,但这些标注的语料往往都没有严格遵循国家规范,词性划分和标注符号的制定细则也五花八门,而且每家单位的语料都紧紧把握在自己手中,除了北京大学计算语言学研究所标注的汉语分词语料库和综合性语言知识库(俞士汶 2003;俞士汶等 2002),能够共享的大规模高质量语言资源寥寥无几,尤其在国际上的影响力都十分有限,而美国语言数据联盟(Linguistic Data Consortium, 简称 LDC)标注的汉语树库(Chinese Treebank)和制定的词性标注规范被广泛地用于各种国际性技术评测,因此国际上从事汉语自动分词和词性标注研究的学者首先想到的可能是 LDC 发布的语料和标注规范,这在客观上让 LDC 的标准成了一种国际标准。这对中国人来说,不能不说是一种悲哀。中国中文信息学会也曾建立过类似的语言资源共享平台(Chinese LDC),一度发挥过很好的作用,但无论是语言资源的规模和种类,还是语料标注的规范性和

质量,以及整个平台资源的权威性,都还很难称得上国家级语料库。俞士汶教授和他带领的团队标注的汉语分词语料库和综合性语言知识库为中文信息处理研究发挥了重要作用,但从国家语言资源库建设和数据共享机制建立的角度考虑,我们还是做得太少。

再看另外一个例子。为了推动“一带一路”倡议的实施,中国与152个国家、32个国际组织签署了共建“一带一路”合作文件,涉及官方语言就有60多种。除了官方语言,“一带一路”沿线国家还有大量的被广泛使用的非官方语言,这些语言对于国际交流、文化传播和国家安全等都具有非常重要的意义。且不说多语言平行对齐的大规模语料库,但就中外文双语对照的高质量数字化词表又有多少呢?“一带一路”倡议从最初提出至今,已经过去了10多年,我们却拿不出一个国家权威的、基本的、通用的中外文对照词表(暂且不说专用领域的双语对照词表)供教学和研发多语言信息处理技术使用。对于信息技术专家来说,往往有经费也不知道到哪里去购买,即使买到了也不知道质量如何,无奈之下只能自己寻求从事相关语种教学或研究的专家帮忙审核,不仅费时费力,造成大量的人力、财力和时间浪费,更为重要的是,分散的语言资源缺乏权威性和统一性,不可能形成国家标准的资源库。

ChatGPT大模型发布之后,有人对其处理的语言种类做过专门分析,据说它可以处理90多种语言,但在英文上的性能表现明显优于其在中文上的表现。ChatGPT(GPT-3.5)模型训练时使用了2021年9月之前全网的文本数据,其中中文语料的占比仅有0.09905%,而英文为92.64708%,这就不难解释为什么中文回答的准确率偏低了。同时这也表明语言资源是影响大语言模型性能的重要因素。

实际上早在20多年前,张普教授就曾明确指出,国家要像对待人力资源、地矿资源、国土资源、森林资源、水源资源一样对待语言资源,语言资源是国家最重要的信息资源(张普2003)。可到目前为止,这种呼吁似乎仍然只是一种美好的愿望。

3.7 考核机制缺乏合理性

合理的考核机制和评价体系是一个学科沿着正确方向发展和个人学术快速成长的重要保障。然而,目前很多绩效考核机制和评价体系却不够合理、公正。本文不想针对学科评估展开讨论,就个人绩效考核和职称晋升而言,其评价系统的不合理性所暴露出来的弊端遭受质疑和诟病在很多高校里并不罕见。例如,在核心期刊上发表论文的数量多少,已经成为很多高校人文社科类教师职称晋升的重要指标。一位教师无论在课程教学中花费多少时间准备教案、制作课件和辅导学生,在教学过程中表现如何出色,如果没有足够数量的高级别论文,同样职称晋升无望。有些教师即便做了大量出色的科研工作,如双语对照的词表整理、语料收集和标注等,但未必都能够在核心期刊上发表论文。去“四唯”的倡议已经提了几年,但目前来看,这种现象不仅没有有效去除,反而似乎愈演愈烈。

当然,考核机制的问题不仅在语言类学科中存在,理工农医等各学科同样存在。重要的是如何走出这种人为设置的怪圈,尽快建立合理的考评机制,真正有效调动教师和学生的科研积极性,让他们安心从事有用的科研工作。

4. 语言类学科走出困境的六点建议

针对人工智能时代语言类学科如何走出困境这一话题,很多专家都有精辟的论述(敖竹梅

2023;彭青龙 2024;郑永年 2024)。笔者从一个“局外人”的视角,提出如下六点粗浅建议。

4.1 开设人工智能通识课,打通学科壁垒

人工智能技术已经广泛应用于各个学科领域,其通用性和普适性已被普遍接受。因此,无论综合性大学,以理工科为主的大学,或以人文类学科为主的大学,人工智能原理都应成为所有院校的通识课。课程内容可以根据不同学科专业的需要灵活把握,例如,对于语言类学科专业的学生可以简要讲解人工智能问题求解的基本原理和实现方法,通过具体案例说明人工智能技术如何在实际应用中发挥作用,主要目的是让学生掌握分析问题、解决问题的理工科思维方式。同时,对于常用的编程语言(如python等),要让学生掌握基本的编程技巧和程序调试能力。对于大模型工具,要让学生熟练掌握其使用技巧,并了解大模型工具可能存在的问题(如幻觉问题、无依据或依据错误等)。进一步,对于从事语言信息处理相关工作的高年级本科生和研究生,应开设自然语言处理课,让学生熟悉自然语言处理的基本方法和工具,如网络爬虫、文本数据清洗、词语切分、文本分类与聚类、统计分析等基本方法,并通过实践掌握基本的文本处理能力。对于具有理工科背景的研究生,可以在课程内容的广度和深度上适当加深。

4.2 外语专业的学生应选修其他专业课程

掌握一种语言的目的是为了获取信息、交流思想和沟通情感。无论是什么语言,都只是一种工具和手段。从人才培养的目的和学生最终就业的方向看,除了极少数学生可能进行语言学理论研究以外,绝大多数学生都要走出高校或研究所,利用所学的语言作为工具服务于社会各行各业。因此,如果学生所学所述的知识只限于本专业领域,势必影响未来的就业和发展。如本文第3.3节给出的例子,在实际应用中需要的知识和技术往往是跨领域、跨专业的。即便学生不参与信息处理系统的研发,仅仅使用语言从事翻译、交流和写作等工作,如果缺乏相关行业或领域的知识,也必然影响能力的发挥。例如,英语专业的学生毕业后到司法、公安部门工作,如果学生学习过司法、政法、公安等相关的专业课程,必然更容易被用人单位接受,工作起来也会更加得心应手。因此,笔者建议面向外国语专业的本科生或研究生开设多个学科专业的选修课,形成“外语+X”的培养模式,如外语+外交学专业课程(外交学/国际政治/国际公共关系等)、外语+法学专业课程(经济法/诉讼法/民商法等)、外语+金融学专业课程(税收学/公司金融学/国际金融学等)。

4.3 更新科研设备,利用新技术手段建立特色教学平台

当学科壁垒被打通之后,语言学等人文类学科专业的科研设备应得到及时更新,为从事一线科研工作的教师和研究生提供基本的软硬件环境,而这就需要学校和学院主要领导转变思想观念,熟悉科研需求,为有能力、熟悉前沿技术、在一线从事科研工作的教师,尤其是青年教师,建立文理工等多学科融合贯通的科研氛围,让他们享受到AI技术的红利,真正融入AI技术发展的大潮,而不是游离于局外,让AI技术为语言研究赋能,包括理论研究和应用技术研发。与此同时,学校和学院应该针对自己的基础、地域等具体情况打造特色专业,与AI技术团队合作研发实用的有独特优势和特色的教学及实践平台。

4.4 科研选题密切结合国家需求

教学和科研是一个学科发展的“两条腿”,缺一不可。科学研究的目标是解释现象,揭示未知奥秘,提出创新理论,建立新技术和新方法,最终的目的是服务于人类社会。任何一项

研究如果不是奔着这样的目标,都是无用的研究,而应用首先应该考虑的是满足国家需求。近年来,国家科技创新2030——“新一代人工智能”重大项目计划支持的“社会治理与智慧社会科技支撑”等重点或重大项目与人文类学科密切相关,这充分体现了理工等多学科交叉融合,联手解决社会重大需求问题的理念。笔者认为,这种合作理念和科研模式为未来学科建设和人才培养提供了正确的方向指引,这也是语言类学科发展和新文科建设的正确路线。

4.5 建设开放共享的语言资源平台

正如前文所述,建立合理的语言资源管理机制和开放平台对于语言资源保护、专业教学和技术研发至关重要。美国LDC语言资源联盟是一个成功的案例,它作为一个非盈利组织,为计算语言学和自然语言处理领域的发展做出了重要贡献,其组织形式和管理模式值得我们学习和借鉴。笔者认为,如果建设国家语言资源共享平台,完全可以采用灵活多样的组织和管理方式。笔者抛出以下五点设想供读者参考:

(1) 国家语言资源管理平台由国家语委统一组织管理,但具体的平台建设和管理任务可以由不同的单位负责,例如,某类语言资源由某个教育部重点实验室负责建立和管理,该平台建设的质量和水平作为重点实验室考核验收的必要任务,根据考核验收情况对实验室进行相应奖惩。所有语言资源(涉密资源除外)清单应该在公开平台上发布。

(2) 严把质量关。无论是语言资源本身(语言学方面)的质量,还是政治立场、思想意识和社会伦理等方面,都必须严格把关,通过鉴定。

(3) 制定合理的奖励机制,调动同行专家的积极性。建立大规模高质量的语言资源库是持续性的任务,因为语言始终是动态发展的,这就需要众多有责任感和使命感的专家共同努力,一代接一代地坚持做下去。当然,国家需要提供经费支持,但同时还需采用奖励机制,例如,为语料贡献者授予某种名衔或者给予某种成果认定,使其在职称晋升(个人)或学科评估(单位)中获益。

(4) 有偿共享语言资源,让资源贡献者获得经济补偿。对学术界和企业界不同的资源使用者可以采用不同的收费标准,资源贡献者和平台管理单位分别获得收益,既是对语言资源开发者的回报,也有利于资源维护和更新。

(5) 由国家经费资助项目所产生的语言资源必须无条件提供给资源共享和管理平台,并通过质量把关,否则不能结题或结项。

4.6 建立合理的考评机制

合理的考评机制可以有效调动个体和集体的工作积极性,让人在愉悦的心情中主动学习和工作,反之,则会起反作用,使个人和集体的工作积极性遭受打击。如本文第三部分所述,目前在很多单位的各类考评中评价机制的合理性仍饱受诟病。鉴于各单位的情况各不相同,本文不就此问题进行过多讨论,只是推荐分轨道(track)评选“代表作”的评价方式。所谓“分轨道”是指根据被考评人所从事的工作分成教学、科研和行政管理等不同的类别,不同的类别评价的核心内容不同,如教学轨道主要评价教师授课的态度、水平、责任感以及受学生欢迎的程度、教材编写情况等,而科研轨道主要评价教师的科研成果。“代表作”是指不以数量为依据,无论承担多少个项目,发表多少篇论文和出版多少部著作,都不作为考量因素,而只看能够代表和体现一个人或者一个团队最高水平的一至三个成果。在极端情况下,如

果一个人发表几百篇论文,而这些论文在国内外同行中都没有任何影响力,也没有实际应用价值,那么这些论文都是无用的。如果一个人只发表一篇论文,但提出了一个新的理论方法或理念,而被国内外同行广泛接受,甚至推动了学科的发展,那么这篇论文充分体现了作者很高的学术水平,足矣。

5. 结束语

学科建设和人才培养是涉及各方面因素、复杂持久的艰巨任务,绝非靠解决几个问题或实施几条措施就可以在短期内达到目的。笔者作为“门外汉”,只是根据自己狭隘的眼见给出粗浅的建议和思考,权当抛砖引玉。文中指出的一些问题和提出的几点建议都未必合适,尤其涉及国家层面的建议或许过于天真和幼稚,但如果能为某个学校或院系提供一点借鉴和参考,也算笔者没有枉费苦心。如果文中有些表述对读者有所冒犯,笔者愿意真诚地表达歉意,并虚心接受任何批评。

致谢:

感谢邓耀臣教授的邀请,如果没有邓耀臣教授的启发和鼓励,我很难挤出时间来梳理我对本文话题的所思所想,并整理成文。本文初稿完成之后,请杨尔弘教授进行了审阅,并根据她提出的宝贵意见对本文进行了修改和补充。在此向两位教授表示衷心的感谢!

参考文献:

- [1] Tesnière, L. 1959. *Éléments de Syntaxe Structurale*[M]. Paris: Librairie C. Klincksieck.
- [2] Chomsky, N. 1957. *Syntactic Structures*[M]. Berlin: Mouton & Co.
- [3] Yingve, V. 1957. Framework for syntactic translation[J]. *Mechanical Translation*, (3): 59-65.
- [4] 敖竹梅. 2023. 人工智能来袭,英语教师“遍地”,英专毕业生如何自救? [OL]. https://m.thepaper.cn/news-Detail_forward_22382128, 引用日期:2024-08-02.
- [5] 冯志伟. 2018. 信息时代需要文理兼通的语言学家[N]. 光明日报, 2018-10-21.
- [6] 刘开瑛. 2000. 中文文本自动分词和标注[M]. 北京:商务印书馆.
- [7] 刘源 谭强 沈旭昆. 1994. 信息处理用现代汉语分词规范及自动分词方法[M]. 北京:清华大学出版社.
- [8] 彭青龙. 2024. 危局·变局·格局:人工智能时代外国语言文学学科和专业的困境与出路[J]. *外语与外语教学*, (3): 20-30.
- [9] 俞士汶. 2003. 语料库与综合型语言知识库的建设[A]. 徐波等编. 中文信息处理若干重要问题[C]. 北京:科学出版社.
- [10] 俞士汶 段慧明 朱学锋 孙斌. 2002. 北京大学现代汉语语料库基本加工规范[J]. *中文信息学报*, (5): 49-64.
- [11] 张普. 2003. 关于汉语语料库的建设与发展问题的思考[A]. 徐波等编. 中文信息处理若干重要问题[C]. 北京:科学出版社.
- [12] 郑永年. 2024. 教育出来的学生学无所用,这是最大的危机[OL]. https://mp.weixin.qq.com/s?__biz=MzA4Nzk2NzEwNA==&mid=2651905385&idx=1&sn=018f813a6189273912759fc77ef90c27&chksm=8ac7073dce7dd05d02e8f930fcf0eec08bcb667ebd10d7688efec60f20fd32900c40360b751e&scene=27, 引用日期:2024-08-16.
- [13] 宗成庆. 2013. 统计自然语言处理(第2版)[M]. 北京:清华大学出版社.

ABSTRACTS

On the Development of Language Disciplines and Talent Cultivation From the Perspective of Artificial Intelligence, p.1. *ZONG Chengqing*

In recent years, the rapid development of Artificial Intelligence (AI) technologies, represented by Large Language Models, has brought great impact on the enrollment, teaching and research of language disciplines and even the entire field of humanities. Foreign language majors have borne the brunt of this impact and have reached a critical juncture. In this context, what opportunities and challenges are faced by the teaching and research of language disciplines and majors? How to get out of the predicament and rebuild the self-confidence of language disciplines has become a hot topic that has attracted much attention. This paper firstly examines the role of linguistic knowledge and linguists from the perspective of AI development, using the foreign language and literature discipline as an example, then points out the seven problems in foreign language teaching and talent cultivation at present, and finally gives six suggestions for the development of foreign language disciplines and majors.

Key words: Artificial Intelligence (AI); foreign language teaching; language studies; foreign language talent cultivation; development of humanities disciplines

Experience Construction and Function Realization of Counterfactual Utterances, p.12. *LIN Zhengjun*

Counterfactuality concerns situations that oppose reality or subjective perception, and counterfactual thinking involves reverse cognitive processes grounded in factual contexts. Traditional linguistic research on counterfactuality predominantly focuses on the lexico-grammatical dimensions, while this study proposes the concept of a counterfactual utterance, offering a novel avenue for linguistic research on counterfactuality. A counterfactual utterance represents an alternative expression of the real-world experience, underpinned by embodied motivations. The counterfactual space is constructed with the lexical-grammatical resources within language. With these resources, counterfactual utterances are generated through the interaction between situational contexts and encyclopedic knowledge. The semantic interpretation of counterfactual utterances is anchored against their counterparts in the real space, thereby contradicting factual or subjective comprehensions. Consequently, counterfactual utterances fulfill their ideational and interpersonal metafunctions by constructing counterfactual experiences.

Key words: counterfactual utterance; counterfactual space; experience; function

A Study of Embodied-Cognitive Multimodal Interaction on Question-Answer Action, p.22. *LIU Yumei & ZHENG Yuying*

Drawing on Multimodal Interaction Theory and the theories from Embodied-Cognitive Linguistics, this paper constructs an analytical framework for embodied-cognitive multimodal question-answer interaction. Taking the question-answer actions retrieved from the program *Chinese Bridge* as cases, this paper then examines the features of their multimodal resources, interaction-