语言智能:语言人工智能研究的历史新方位

——“语言智能科学”理论与方法论构建(三)

姜孟

(四川外国语大学 语言智能学院(通识教育学院)/语言脑科学研究中心,重庆 400031)

摘 要:寓身于人肉身之内的语言智能是语言自然智能,离身于人的肉身靠人造装置实现的语言智能属于语言人工智能。当今,致力于研究前者的典型学科是语言学,致力于研究后者的学科则有机器翻译、计算语言学、自然语言处理等。本文站在10~30年后的学科未来看历史与现今,透过学科指涉名称上的差异解析其背后的思想理路、主线脉络、方法重心、技术逻辑、历时承继与未来趋向,提出新的主张与判断:语言人工智能研究已经走过了思想乌托邦(前机器翻译)、泛机械(机器翻译)、语言学主导的符号主义(计算语言学)、计算机科学主导的连接主义(自然语言处理)四个历史方位,正在迎来第五个崭新的历史方位——“智能科学主导的机制主义”。新近提出的“语言智能”概念是这一历史方位恰当的代名词。

关键词:语言自然智能;语言人工智能;历史方位;机制主义

中图分类号:H087 **文献标志码**:A **文章编号**:1674-6414(2024)04-0060-21

0 引言

以人的碳基身体为基础生发的语言能力,寓于人的身体之内,出自人类生命体的天然与本能,彰显的是人类语言能力的自然维度,可被称作“语言自然智能”。凭借人造装置在人的碳基身体之外实现的人的语言能力,是对人类语言能力的模仿、延伸与扩展,彰显的是人类语言能力的人工维度,可称作“语言人工智能”。当前,致力于研究前者的公认学科是语言学,而致力于研究后者的学科则有好几个(至少从名称术语来看是如此),如机器翻译、计算语言学、自然语言处理、自然语言理解等。最近,又出现了“语言智能处理”(黄河燕 等, 2020)、“机器语言能力研究”(耿立波 等, 2014)、“语言智能”(周建设 等, 2017)等

收稿日期:2023-10-15

基金项目:四川外国语大学2021年哲学社会科学重大项目“语言智能新文科体系构建及其范式创新研究”(sisuzd202104)的部分成果

作者简介:姜孟,男,四川外国语大学语言智能学院(通识教育学院)教授,博士,博士生导师,主要从事认知神经语言学、语言病理学、语言智能与语言脑机接口研究。

引用格式:姜孟. 语言智能:语言人工智能研究的历史新方位——“语言智能科学”理论与方法论构建(三)[J]. 外国语文,2024(4):60-80.

概念术语。这些带有学科指向的概念术语之间究竟有何区别与联系,映照了怎样的历史、现实与未来取向?从1949年瓦伦·韦弗(Warren Weaver)正式提出“机器翻译”的思想算起(见后文论述),语言人工智能研究已经走过了75年的历程。这75年背后究竟有何发展逻辑与轨迹可循,这一学科前进到了什么历史方位,正走向哪儿,这些概念术语各自有何意蕴?带着这些疑问,本文试图站在10年、20年乃至30年以后,语言人工智能研究已经十分发达成熟的学科位点,回望历史,检视当下,着力探寻这些问题的答案,以期更好顺应、助力本学科的发展。

1　人类语言能力研究的两个维度

人工智能之父艾伦·图灵(Alan Turing)1948年对人类智能给出了独特的理解。他认为,“智能”可分为“Embodied Intelligence”与“Disembodied Intelligence”(李德毅, 2023/2024)。前者指与人的身体紧密结合、存在在于人身体之内的智能,即“具身智能”;后者指脱离人身体、存在于人身体之外的智能,即“离身智能”。具身智能以人自然的碳基身体为依托,折射的是人类数百万年生物进化的成就,因此属于“自然智能”的范畴。离身智能是借助一定的装置或设备来实现对人的智能的模仿、复制、延伸或扩展,超越了人的肉身,人造的特征很突出,是一种“人工智能”。

对世界万物做“自然—人工”划分的思想由来已久。早在2 000多年前,古希腊哲学家亚里士多德就从他的“四因说”(质料因、动力因、形式因、目的因)出发区分了“自然物”与“人工物”。他认为,凡是自然物,都具有内在的生成和运动根源;凡是人工物,其生成和运动的原因都不在于自身,而在于人的目的(亚里士多德, 1982)。把这一思想应用于“智能”,也就顺理成章区分了“自然智能”与“人工智能”。从“智能普存论”的立场来看(李珍, 2020: 46),自然智能是自然界盲目选择过程中的特定阶段所产生的生物智能,是生物进化的产物;人工智能是在非生物基础上的智能创造,是人类有目的地制造出来的人工物,源于对自然智能的模仿与复制。借用塞尔“内在意向性”(intrinsic intentionality)与“派生意向性”(derived intentionality)的二分思想(塞尔, 2006: 69),自然智能是具有内在意向性的智能,而人工智能则只具有派生意向性的智能。事实上,正是在图灵“具身—离身”智能二分思想的影响下,信息论创始人克劳德·艾尔伍德·香农(Claude Elwood Shannon)等人在1956年达特茅斯会议上正式提出了当今的“人工智能”(Artificial Intelligence)概念。李德毅(2024: 249)指出:“把人类智能释放到体外,成为离开生命体的存在,即人工智能,包括机器智能。”由此,人工智能与自然智能相依相存,二者合为完整的“智能”。

基于以上智能二分的思想,对人类语言能力的研究也可从两个方面来进行。一是着眼于人类语言能力的自然维度,把语言能力看作是人类生命体在自然进化基础上的先天、后

天交互成就,探究其本质、机制与特征;另一种就是着眼于人类语言能力的人工维度,脱离人的碳基身体,通过构建一定的物件与装置,来再现、再造、扩展人的语言能力。前者属于“语言自然智能研究”,后者属于“语言人工智能研究”。当今的语言学及诸多交叉边缘学科,如神经语言学、心理语言学、计量语言学、语料库语言学等,都属于致力于语言自然智能研究的学科;“机器翻译”“计算语言学”“自然语言处理”等概念术语所指称的学科,则都属于致力于语言人工智能研究的学科①。

基于此,下文用“语言人工智能研究”这一术语来统称、泛指国内外一切致力于把人类语言能力“释放到体外”,让其成为“离开生命体的存在”的研究(李德毅, 2024),以避开现有概念术语在内容指称上的交错混叠。

2 现今语言人工智能研究的四个历史方位

2.1 第一个历史方位:思想乌托邦(前机器翻译)时期(古希腊—1949)

人类很早就萌生了让机器拥有人的部分或全部语言能力的思想。最直接、最实际的想法就是制造机械装置,代替人进行跨语言翻译活动。这可以说就是最早的语言人工智能研究。

远在古希腊时期,就有人提出利用机械装置进行语言翻译的想法(李沐 等,2019: 2-3)。17 世纪,笛卡尔(Descartes)和莱布尼茨(Leibniz)提出使用机器词典来实现语言翻译。17 世纪中叶,“普遍语言”运动提出设计一种无歧义的“中介语”来实现世界各语言之间的对应转换。1903 年,古图拉特(Gouturat)和洛(Leau)采用德国学者里格(Rieger)提出的一种数字语法,基于词典的辅助,利用机械装置将一种语言翻译成多种语言。同年,他们出版了《通用语言的历史》一书,并首次使用了“机器翻译”的术语。20 世纪 30 年代初,法国工程师阿尔楚尼(Artsouni)提出基于存储装置进行语言翻译的想法。1933 年,苏联发明家彼得·特洛延斯基(Peter Troyanskii)提出使用双语字典和语言间的语法角色完成翻译的想法。1947 年 3 月 4 日,信息论先驱、美国科学家、“机器翻译之父”沃伦·韦弗(Warren Weaver)写信给控制论之父诺伯特·维纳(Norbert Wiener),探讨利用机器进行语言翻译的可能性。他认为,由于语言中存在语义困难的问题,机器翻译的质量很难达到“雅”,但科技文献的翻译很可能达到“信”。1949 年 7 月 15 日,他在题为《翻译》的备忘录中正式提出了机器翻译的思想,其核心要点包括(孙茂松, 2016: 13-14):

(1)上下文与(词汇)语义解歧:引出了语言的统计语义问题,与后来的马尔科夫语言模型相对应。

(2)语言与逻辑:书面文本是逻辑性质的表达,可进行结构化的句法语义分析,具有形式上可解特征;语言中的风格感受、情感内容等非逻辑元素,不具备形式可解特征,很难被计算

① 参阅李宇明(2023: 4)提出的“人工语言智能”概念。

机处理。

(3)翻译的编码与解码性质:从密码学角度,一本中文书可视作是一本英文书的“编码”,其翻译过程则是“解码”。

(4)人类通信的共同基础:普遍语言(也称“语言的逻辑结构”):与后来学者提出的机器翻译“中间语言”思路一脉相承。

从古希腊到1949年韦弗机器翻译思想的提出,这一时期长达数百上千年,堪称语言人工智能研究的第一个历史方位。这一方位的整体成就在于:萌生了让语言能力“附身”机器的想法,产生了从机器翻译入手来突破语言人工智能的思想与路径,但总体上还处于思想启蒙、前路茫茫的阶段,离真正的“把思想变现”还有相当距离,可以称作“思想乌托邦”时期。

2.2 第二个历史方位:泛机械(机器翻译)时期(1949—1966)

韦弗的备忘录从思想上对机器翻译进行了启蒙,引发了大量机器翻译研究实践。最早开展机器翻译研究的有美国的麻省理工学院、乔治城大学和IBM(国际商业机器公司)以及苏联的列宁格勒大学、英国的剑桥大学等。美国主要进行了俄英机器翻译试验,苏联则主要进行了英俄和法俄机器翻译试验。1952年在美国麻省理工学院(MIT)召开了第一次机器翻译会议,1954年出版了第一本《机器翻译》(*Machine Translation*)杂志。1954年1月7日,美国乔治敦大学和IBM公司使用IBM-701计算机开发了世界上第一个机器翻译原型系统,成功地将60多句俄语自动翻译成英语。尽管该系统还非常简单,仅包含六个语法规则和250个词,但这一事件成了机器翻译史上的一个里程碑事件。

我国早在1956年便将机器翻译列入国家科学工作发展规划。1958年8月,中国科学院计算技术研究所牵头成立了机器翻译研究组,与语言研究所合作开展俄汉机器翻译研究。1959年,我国在自制通用电子计算机上进行的俄汉机器翻译试验获得成功(刘涌泉, 1963)。

这一时期,研究者先是对机器翻译普遍持高度期待、高度乐观的态度,但随着研究与实践的进展,人们发现机器翻译的质量与期望相去甚远,“语义障碍”(semantic barrier)构成了一个绕不过的难题。1960年,以色列著名的哲学家、数学家和语言学家耶和书亚·巴尔-希勒尔(Yehoshua Bar-Hillel)发表长文指出,由于语义歧义的存在,通用的高质量全自动机器翻译在理论上是不可能的。1966年11月,美国科学院自动语言处理顾问委员会(Automatic Language Processing Advisory Committee,简称:ALPAC)和美国国家研究理事会,发布了题为《语言与机器:翻译和语言学视角下的计算机》的著名报告(也称ALPAC报告)。报告正文30多页,附件90页,对机器翻译作出了基本负面的评价(因此也被称为“黑皮书报告”),其要点包括:第一,机器翻译遇到了难以克服的语义障碍问题,全自动机器翻译在可预期的将来不可能达到与人的翻译相比更为快速、更为经济、质量更优的水平,在目前还没有多少理由给予机器翻译以大力支持;第二,应该加强对机器辅助翻译和计算语言学(Computational Linguistics)

的研究与支持(冯志伟,2019; 孙茂松,等, 2016)。这成了语言人工智能研究的一个转折点。

现在回看,这是一个很特别的时期。人类初涉语言人工智能实践研究,在人的诸种语言能力中好不容易选定“翻译”作为突破口,也好不容易有了当时看来计算能力和存储能力都十分强大的新机械——“计算机”,似乎对语言(翻译)能力的机器实现不再是什么难事。于是,研究者们开展了如火如荼的机器翻译系统研发工作。在实践中,尽管所依靠的主要是计算机而非其他什么机械或装置,但在他们心中计算机只不过是机械或机器的化身,是一种最典型、最现实、最称心上手的机械装置。

这一时期的语言人工智能研究有两个特点。一是在研究方法论和核心技术路线上“初生牛犊不怕虎”。研究者们低估了语言的复杂性,以为只要由人来编制一些词法、句法规则(例如,美国乔治城大学的自动翻译系统 GAT 就配置了词法层、组合层和句法层)或设计一种“中间语言(interlingua)”(如欧洲和苏联的所开展的机器翻译研究)就可以彻底实现机器翻译的目标(即走的是上述韦弗机器翻译思想②与④的技术路线)。这一时期的另一个特点是对技术的基础性、支撑性作用认识不足,有些简单天真。研究者们折服于当时计算机令人耳目一新的计算能力和存储能力,以为从此可以万事大吉,借此让机器拥有语言能力的梦想变为现实,没有预料到机器翻译以至整个语言人工智能研究对技术的要求太高,今后都还要牢牢地受限、受制于计算机技术及其高阶人工智能技术。也正是这两个特点,决定了这一历史方位是“泛机械(机器翻译)时期”,即语言人工智能研究找到了机械技术做靠山,有了“机巧”,但尚没有傍依“学科”,是无学科支撑的时期。

2.3 第三个历史方位:语言学主导的符号主义/理性主义(计算语言学)时期(1966—1980)

1966 年美国 ALPAC 报告的发布,使“机器翻译”这一当时最亮眼的语言人工智能工程的资金被大量消减或中断。面对困难与困境,语言人工智能研究被迫另寻他途。好在 ALPAC 报告在基本否定机器翻译可能性和前景的同时,也给出了研究者可以和应该用力的方向。ALPAC 报告的主起草人之一、美国语言学家戴维·海斯(David Hays)在报告中给出建议:在放弃机器翻译这个短期工程项目的时候,仍有必要加强语言和自然语言计算机处理的基础理论研究,应当把原来用于机器翻译研制的经费使用到自然语言处理的基础理论方面(冯志伟,2019)。所谓的“自然语言处理的基础理论方面”正是指语言计算方面的研究,海斯把它命名为“计算语言学”(Computational linguistics)。这是“计算语言学”概念首次在官方报告中使用。1967 年,海斯又出版专著《计算语言学导论》(*Introduction of Computational Linguistics*),宣告了计算语言学时代的正式到来。机器翻译研究走上了先做好语言和语言计算基础理论研究再图将来的迂回发展之路。

从今天或未来某个更远的位点来看,语言人工智能研究将主要受两大力量助推。一是对语言自然智能本质的揭示,也就是语言学这一“内容性科学”在研究上的突破带给语言人工

智能的推力;再就是以算法、算力、算料(数据)为核心的计算机技术的发展和提升带给语言人工智能的推力。前者是一种思想内容性推力,后者则主要是一种技术方法性推力。在前一历史方位,研究者刚刚迎来用机械或机器代替人做跨语言翻译的喜悦,几百年梦想成真,真是惊天大喜,还不及思考语言人工智能研究今后将会怎样“道阻且长”,1966 年发布的 ALPAC 报告直接结束了这一短暂的兴奋,蓦然间“穿越”到了困难重重的现实。置身困境,一个最自然、最简便的脱困之法便是傍依语言学,寻求语言学的帮助。于是,语言人工智能研究走向了寻求“语言学靠山”的历史方位,希冀从语言本体特征与规律的研究中寻求机器翻译上的突破。

从实际研究来看,在 1966 年以后的十多年间(1966—1980),无论是机器翻译还是整个计算语言学研究,走的也确实都是聚焦“从语言学角度,分析自然语言的词法、句法等结构信息,并通过总结这些结构之间的规则,达到处理和使用自然语言的目的”的路子(毕然 等, 2022:243)。这实际上是美国语言学家乔姆斯基和他所提出生成式文法所开创的“符号主义”的路子,也就是文献中常提及的“基于规则的”研究路径。因此,语言人工智能研究的这一历史方位可以称作“语言学主导的符号主义(计算语言学)”时期。

从机器翻译到计算语言学的“转拐”发生,早有端倪。早在 1962 年,美国就成立“机器翻译和计算语言学学会”(Association for Machine Translation and Computational Linguistics, 简称:AMTCL)。这也是“计算语言学”作为一个学术概念首次被提出,尽管当时获得的影响力和认可度还有限。紧接着,1965 年由 AMTCL 主办的《机器翻译》(*Machine Translation*)杂志更名为《机器翻译和计算语言学》(*Machine Translation and Computational Linguistics*),以使杂志与主办的学会名字完全一致。需要特别指出的是,这次更名,虽然在杂志的封面上首次出现了“Computational Linguistics”字眼,但“and Computational Linguistics”这三个单词是用特别小号的字母排印的(冯志伟, 2011)。这刚好从一个侧面说明了,计算语言学这个今天影响很大的学科产生自人们对机器翻译研究的失望、无奈与信心不足,有着很强的迫不得已、另寻他途的意味。即便当时人们对“Computational Linguistics”能否成为一门真正独立的学科还没有充分的把握(冯志伟, 2011),还不敢公开拿它与“Machine Translation”分庭抗礼,但也要将其像救命稻草一样抓住,视其为备案与退路。

实际上,“机器翻译”怎么看也都只能是人类探索语言人工智能的一个小瞬间,是一个历史“小不点”。前进的步伐注定要超越机器翻译的小天地,走向更大的海阔天空。到了 1968 年,“机器翻译和计算语言学学会”(AMTCL)干脆把“Machine Translation”这两个词删除,更名为“计算语言学学会”(Association for Computational Linguistics, 简称:ACL),并一直沿用至今。另外,在 ALPAC 报告发布前的 1965 年,在美国纽约还成立了单独以“Computational Linguistics”冠名的国际计算语言学委员会(International Committe of Computational Linguistics,

简称:ICCL)。该学会每两年召开一次国际会议,会议名称为“计算语言学国际会议”(International Conference on Computational Linguistics,简称:COLING)。与此同时,美国出版了学术季刊《美国计算语言学杂志》(*American Journal of Computational Linguistics*),后改名为《国际计算语言学杂志》(*International Journal of Computational Linguistics*)。

“计算语言学”替代“机器翻译”成为语言人工智能研究的历史新方位,非一朝一夕之事。语言计算研究的思想由来已久,远端可以追溯到 19 世纪俄国数学家布依考夫斯基(B. Buljakovski)、英国数学家德莫尔芬(A. DeMorgen)、瑞士语言学家德·索绪尔(De Saussure)、德国学者凯定(F. W. Kaeding)等人的数理语言学思想,近端可以追溯到 20 世纪四五十年代信息论的创始人香农(Shannon)、人工智能之父阿兰·图灵和转换生成语言学创始人诺·乔姆斯基等人的“语言熵”“图灵测试”“上下文无关语法”等思想(冯志伟, 2011)。

应当指出的是,我们用“符号主义”来命名这一历史方位,并不是说“计算语言学”所指称的学科从头至尾,直到今日,走的都只是语言学主导的符号主义的研究路子,没有越雷池一步。而仅是说,计算语言学就其诞生时所承继的历史方位与所承载的初心与取向来说,是“符号主义”的。它代表着语言人工智能研究在遭遇困难时寻觅语言学靠山,走符号主义研究路子以获得出路的尝试与努力。实际上,从 20 世纪 90 年代初开始,计算语言学就逐步走上了以概率为基础、以“连接主义”为导向的研究路子。另外,从机器翻译进入计算语言学历史方位后,并不是说机器翻译研究从此停止了,而是说计算语言学从此成了语言人工智能研究学科的代表,机器翻译被收归其下,成为其一个分支领域,丢掉了原先代表这一学科发展某个特定历史方位的资格与地位。换个角度看,过去机器翻译只是着眼于人类“听说读写译”五种能力中“译”能力的机器实现问题,现在计算语言学则从计算机处理自然语言所必须经过的“形式化、算法化、程序化和实用化”四个过程的角度来看待语言(冯志伟, 1996/2011/2019),面更广,格局更大。

2.4 第四个历史方位:计算机科学主导的连接主义(自然语言处理)时期(1980 年至今)

在符号主义时期,基于规则的语言人工智能研究获得了长足发展,但随着人们越来越多地关注工程化、实用化的解决问题的方法,这一研究路径日渐遭遇了困难。就拿机器翻译来说,人工确定的翻译规则越来越复杂,规模库也越来越大,但数量却有限,对千变万化的复杂语言现象的解释力难以继续提高,译文的准确率也无法持续改善,语言人工智能研究又来到了一个新的历史拐点。这一次,研究者寻找的是以概率统计为核心特征的“技术”靠山,傍依的是快速发展的计算机科学,走的是“连接主义”的研究路子(毕然 等, 2022: 243),最能反映这一发展现实的学科名称是“自然语言处理”。

从 20 世纪 80 年代开始,一些研究者就开启了离开符号主义而另走他途、再寻前程之旅。1980 年马丁(Martin Kay)提出了翻译记忆(translation memory, 简称:TM)的方法,尝试从已经

翻译好的文档中找出相似部分来帮助新的翻译。1984 年长尾真(Makoto Nagao)提出了基于实例的机器翻译方法(example-based machine translation, 简称:EBMT),着手从实例库中提取翻译知识,通过增、删、改、替换等操作完成翻译(李沐 等, 2019: 4)。这些研究所尝试的都是全新的基于数据驱动的机器翻译方法,堪称语言人工智能研究走向第四个历史方位的先声。

与此同时,计算机硬件技术迅猛发展,存储容量快速扩大,运算速度日益提升,统计机器学习的新理论、新方法不断涌现,语料库技术也日臻成熟,使得很多原来无法实现的复杂问题现在都可以借助“硬件技术+统计模型+语料库”很容易地实现。1990 年,在芬兰赫尔辛基召开了第 13 届国际计算语言学大会,提出了处理大规模真实文本的战略任务,开启了语言计算的一个历史新阶段——基于大规模语料库的统计自然语言处理。在此潮流的带动下,1993 年,美国 IBM 研究院发表论文“统计机器翻译的数学理论:参数估计”(The Mathematic of Statistical Machine Translation: Parameter Estimation)(Brown et al., 1993),提出了 IBM Model 1~5。1999 年,美国约翰·霍普金斯大学(The Johns Hopkins University)发布了 GIZA 软件包,把 IBM Model 1~5 变为了现实。随后,更加复杂的 IBM Model 6、更加优化的软件包 GIZA++ 都先后发布。IBM 统计机器翻译模型得到广泛使用。该模型几乎完全依赖大规模双语语料库,通过词对齐、短语对齐等手段,来自动构造统计机器翻译系统。较之基于规则的机器翻译系统,统计机器翻译系统的性能显著提升,而且很容易实现数十种语言之间的翻译。由于统计机器翻译模型与具体语种无关,不再需要“规则集”,设计者可以完全不懂相关的语言。机器翻译研究回归到了前述韦弗翻译思想之(1)和(3),走上了离语言学研究越来越远的道路(孙茂松 等, 2016: 16)。为此,著名的机器翻译学者、谷哥翻译(Google Translate)的设计者弗兰茨·约瑟夫曾信心满怀地声称:“只要给我充分的并行语言数据,那么,对于任何两种语言,我就可以在几小时之内给你构造出一个机器翻译系统。”

2006 年,深度神经网络反向传播算法被提出。2014 年前后,随着深度学习技术在语音、图像领域的研究取得成功,深度学习方法也开始应用于语言人工智能研究。在机器翻译领域,出现了“神经机器翻译”(Neural Machine Translation, 简称:NMT)模型。该模型采用基于深度神经网络的方法来构造机器翻译系统(Bahdanau et al., 2014; Sutskever et al., 2014),其架构由编码器和解码器两部组成。首先由编码器把源语言的句子表示为词向量(word vector),形成句子的分布式,然后利用解码器依次生成目标语言的单词序列,直到生成目标语言的整个句子为止。神经机器翻译采用的是端到端(end to end)的计算过程,由于其内部是由基于词向量的数值计算构成的,难以从语言学的角度解释中间过程的计算机制,翻译成为一个黑箱操作过程(李沐 等, 2019)。相比统计机器翻译,神经机器翻译具有更加广泛的一般性,更加远离具体语言的知识。机器翻译走上了与语言学研究几乎彻底分道扬镳的道路。在技术上,神经机器翻译具有更大的“颠覆性”,它使机器翻译的质量再次迅猛提升。当然,

它涉及的计算量也更大,更加依赖计算能力(“算力”),需要借助特殊计算设备(如 GPU)才能高效地实施参数训练。进入 20 世纪 90 年代以后,整个语言人工智能领域都深深地打上了概率统计的烙印。正如冯志伟(2011:13)所指出的那样,“概率和数据驱动的方法几乎成了计算语言学的标准方法。句法剖析、词类标注、参照消解、话语处理、机器翻译的算法全都开始引入概率并且采用从语音识别和信息检索中借过来的基于概率和数据驱动的评测方法。”

这一历史方位实际上是语言学主导的符号主义研究方法在其红利快要耗尽之时,语言人工智能研究寻觅到的新质增长点与靠山。相比前一方位,这一方位的研究几乎完全依靠概率统计、深度神经网络以及计算机硬件技术,研究者所积累的语言学和计算语言学专业知识变得几乎完全不起作用,语言学被边缘化,计算机科学成了主宰,“连接主义”成了新的标示与旗号。

与上一历史方位不一样,在这一历史方位转拐之际,未出现像“机器翻译”“计算语言学”那样,可用以标示转拐现实的崭新专门概念术语。就现有概念术语来看,“自然语言处理”(Natural Language Processing, 简称:NLP)很好体现了由技术主导的这一历史方位的特征(笔者未查到自然语言处理概念提出的具体时间)。且看一些学者对“自然语言处理”的定义。冯志伟(1996)认为,自然语言处理就是利用计算机为工具,对人类特有的书面形式和口头形式的自然语言的信息进行各种类型的处理和加工的技术。他直接将这一概念定义为一种“处理和加工的技术”。毕然等(2022)认为,自然语言处理是一门融语言学、计算机科学和数学于一体的科学,主要研究人与计算机之间使用自然语言进行有效通信的各种理论和方法,不仅研究语言学,还研究能高效实现自然语言理解和自然语言生成的计算机系统,尤其是其中的软件系统。在他们看来,“自然语言处理”中的“处理”二字的含义体现在,“计算机以用户的自然语言数据为输入,在其内部通过定义的算法进行加工、计算等系列操作后(用以模拟人类对自然语言的理解),再返回用户所期望的结果”(毕然 等, 2022:240)。可以说,这二字直指计算机算法技术。他们还特别强调,自然语言处理是计算机科学的一部分。宗成庆(2019:4)也明确指出,相比计算语言学,自然语言处理似乎“包含的语言工程和应用系统实现方面的含义似乎更多一些”。因此,我们认为这一历史方位的代名词就是“自然语言处理”。

同样,在从“计算语言学”进入“自然语言处理”历史方位后,并不是说计算语言学所代表的符号主义路向的研究即告结束与消失,而是说它让位于自然语言处理所代表的连接主义的研究路向,不再居于主流、支配地位,失去了代表语言人工智能研究某一特定历史方位的资格与地位。实际上,在语言人工智能研究实践中,“计算语言学”与“自然语言处理”都一直被用作整个学科领域的代名标签,两者并驾齐驱,不分伯仲。本文对它们做区分,认为它们各自代表语言人工智能研究的一个历史方位,是从两个术语各自的“本义”与“初心”来说的。黄河燕(2020:1)等明确指出“自然语言处理早期也被称作计算语言学”。言下之意,计算语言学

与自然语言处理确实各自代表了语言人工智能研究的不同阶段或时期。

3　第五个历史方位:智能科学主导的机制主义(语言智能)时期(今天—)

3.1 从学科之名来看第五个历史方位

语言人工智能研究走过了四个历史方位,从“思想乌托邦”的前机器翻译时期到“泛机械”的机器翻译时期,再到“语言学主导的符号主义”计算语言学时期,直到“计算机科学主导的连接主义”自然语言处理时期。当下,语言人工智能研究正在迎来第五个历史方位,即“智能科学主导的机制主义(语言智能)”时期。这一最新态势可以从该学科“名”与“实”的现状中洞悉与察知。

从“名”来看,语言人工智能研究兴起于 1950 年前后(毕然 等, 2022: 242),至今 70 多年。但长期以来,该领域一直缺少一个有统摄力的学科名称,多个概念术语交叠瓜葛,纷争抵牾,各自为政,但最近出现了向智能及智能科学相关名称聚靠的趋势。

在语言人工智能研究的四个历史方位中,有三个比较有影响的术语。第一个是“机器翻译”。这一概念由古图拉特(Gouturat)和洛(Leau)于 1903 年提出(如前所述),距今 120 余年,它比今天的“人工智能”(Artificial Intelligence)概念都早了 50 余年(按 1956 年达特茅斯会议正式提出此概念算),历史厚重。“机器翻译”是指“使用机器(计算机)自动地将一种自然语言(源语言)的语句转化为相同含义的另一种自然语言(目标语言)的语句的过程”(李沐等, 2019: 2)。显然,这一概念突显的是人的听、说、读、写、译多项语言智能中的“译”的智能,仅关注人的翻译智能的机器实现(即模仿、延伸与扩展),只涵盖了人类语言自然智能之五分之一,无力承担指称语言人工智能这一被誉为“人工智能皇冠上的明珠”的前沿交叉学科全域的重任。

第二个术语是“计算语言学”。该术语于 1966 年由美国科学院在 ALPAC 报告中正式提出(见前文)。该术语的本义与初心体现在以下两个简明的定义中:“计算语言学是利用数字计算机进行的语言分析”“计算语言学是语言学的一个分支,专指利用电子计算机进行语言研究”(宗成庆, 2019: 3)。这一术语在早期,可以说仅是强调计算机对于自然语言分析与统计的辅助工具作用,还没有把计算机上升到“可以提供主动服务,能够帮助人类达到对话、翻译、检索等若干目的的智能工具”的高度(宗成庆, 2019: 3)。只是到了后来,语言人工智能的研究实践远远超越了计算语言学这一概念的原初内涵与范围,但却又找不到一个相称、合适的术语之时,这一概念才被迫用作统摄语言人工智能研究全域的一个术语,颇有些“不欲戴王冠却得承其重”的意味。这可见于后来一些学者对“计算语言学”所做的理解与解释中。例如,克里斯特尔(2000)在《现代语言学词典》中给出的定义是:语言学的一个分支,用计算技术和概念来阐述语言学和语音学问题。所开发的领域包括自然语言处理、言语合成、言语

识别、自动翻译、编制词语索引、语法的检测,以及许多需要统计分析的领域。宗成庆(2019:4)的定义与解释也体现了这一点:“计算语言学实际上包括以语音为主要研究对象的语音学基础及其语音处理技术研究和以词汇、句子、话语或语篇及其词法、句法、语义和语用等相关信息为主要研究对象的处理技术研究。”可见,“计算语言学”原本就是一个偏重语言技术方法研究方面的概念,并非指称统揽语言人工智能昨天、今天和明天广阔研究实践的理想术语。

第三个术语是“自然语言处理”(Natural Language Processing,简称:NPL),也称“自然语言理解”(Natural Language Understanding,简称:NLU)。该术语源自机器翻译,但严格说它算不上一个专业术语,而仅是对某种研究对象与内容的描述。正如莫宏伟等(2020:26)所指出的,从机器翻译开始,人工智能领域发展出了自然语言处理“这一研究内容”。根据冯志伟(1996)的定义,自然语言处理就是利用计算机为工具,对人类特有的书面形式和口头形式的自然语言的信息进行各种类型的处理和加工的技术。因此,在本义上,“自然语言处理”是一个技术至上的概念。用宗成庆(2019:4)的话来说,自然语言处理似乎“包含的语言工程和应用系统实现方面的含义似乎更多一些”。然而,语言十分复杂,当今人工智能学科对语言开展研究的范围又十分广阔,涉及人类语言机器实现的任何技术都隐含着当今自然语言处理或计算语言学的问题,广泛牵涉其他学科领域的方法与技术,如信息检索、舆情分析、文字识别、社交网络、社会计算、情感计算、语言教学、口语考试自动评分等(宗成庆,2019)。因而,自然语言处理天生就不限于自然语言处理。在现实中,同样由于缺乏更恰当的概念术语,与“计算语言学”概念一样,“自然语言处理”也被揠苗助长般地用于指称当今语言人工智能研究全域。美国计算机科学家马纳瑞斯(Bill Manaris)(1999)的定义很好地诠释了这一点:“自然语言处理是研究人与人交际中以及人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力和语言应用的模型,建立计算框架来实现这样的语言模型,提出相应的方法来不断地完善这样的语言模型,根据这样的语言模型设计各种实用系统,并探讨这些实用系统的评测技术。”(宗成庆,2019:5)我国学者刘涌泉在《大百科全书》中所给的解释也展示了这一概念当下所承载的广阔学科内涵。他认为,自然语言处理是人工智能研究的主要内容,“即利用电子计算机等工具对人类所特有的语言信息(包括口语信息和文字信息)进行各种加工,并建立各种类型的人—机—人系统。自然语言理解是其核心,其中包括语音和语符的自动识别以及语音的自动合成”(宗成庆,2019:3)。在现实中,自然语言处理所涵盖的研究内容与范围已经十分广泛,仅从应用目的的角度,它至少包括:机器翻译、自动文摘、信息检索、文档分类、问答系统、信息过滤、信息抽取、文本挖掘/数据挖掘、舆情分析、隐喻计算、文字编辑与自动校对、作为自动评分、光读字符识别、语音识别、文语转换、说话人识别/认证/验证等。严格意义上,它还应包括语音技术,即语音识别、语音合成和说话人识别等(宗成庆,2019)。黄河燕等(2020)也认为,自然语言处理的研究对象几乎涉及语言学研究的所有对象:

语音、形态、语法(句法)、语义、语用,研究内容包括针对这些对象的自动分析方法与技术,如词法分析、句法分析、语义分析等。

可见,学界当前的两个核心概念“自然语言处理”与“计算语言学”都算不上是指称语言人工智能研究全域的理想术语。况且,两者间还存在一定的抵牾。自然语言处理曾一度被视作计算语言学的分支领域,是其下位概念。例如,《现代语言学词典》(克里斯特尔, 2002)就把自然语言处理视为与言语合成、言语识别、自动翻译、编制词语索引、语法的检测、文本考释等并列的计算语言学的开发领域之一。随着近年来自然语言处理被视作语言人工智能研究全域的代名词,它已经超越作为计算语言学范畴内的一个研究分支的地位,而被看作“基本上是处于同一个层次上的概念”(宗成庆, 2019: 4)。两者(连同“自然语言理解”)很多时候都被视作同一个概念,至少在其外延上不再细究其差异。用宗成庆(2019: 4)的话来说:“在很多情况下我们很难绝对地区分开‘计算语言学’‘自然语言理解’与‘自然语言处理’三个术语之间到底存在怎样的包含或重叠关系以及各自不同的内涵和外延。”它们被视为是等同的,即自然语言处理就是自然语言理解,也就是计算语言学(刘颖, 2002)。

正是由于语言人工智能研究陷于“名不正”的情势与地位,近年来一些学者提出了好几个新术语,以回应、矫正、补足这一现实,更好满足各自研究的需要。这些术语几乎都是通过对原有术语掐“头”去“尾”、添“智”加“能”的方式创生的,智能化的趋势与倾向明显。

例如,黄河燕等(2020)就提出了“语言智能处理”的概念,以更精准地指称当今被高度智能化、带有显著智能特征的自然语言处理的研究现实与实践。根据她们的界定,语言智能处理主要体现为自然语言处理,是指“利用计算机等工具分析和生成自然语言(包括文本、语音等),从而让计算机‘理解’和‘运用’自然语言。通过自然语言处理的一系列方法和技术,可以让人类通过自然语言的形式与计算机系统进行智能交互”(黄河燕 等, 2020: 1)。他们认为,语言智能处理是人工智能领域的重要研究方向,涉及计算机科学、语言学、逻辑学、数理统计、认知科学等诸多学科,具有显著的跨学科特色。张雄伟等(2020)提出了“智能语音处理”(Intelligent Speech Processing)的概念。智能语音处理与经典语音处理相对。经典语音处理方法以语音产生和语音感知为研究重点,以语音短时平稳和线性模型为基本假设,通过语音特征参数提取和数字信号处理的手段来实现语音处理的目标。但近十年来,随着人工智能技术快速发展,机器学习新技术、新算法不断涌现,尤其是新型神经网络和深度学习技术的出现,语音处理研究获得了新的研究方法与手段,智能语音处理应运而生。后者的特点是从大量语音数据中学习和发现其中蕴含的规律,以有效解决经典语音处理难以解决的非线性问题,提升传统语音应用系统的性能,为语音新应用提供性能更优的解决方案。他们认为,在广义上,“在语音处理算法或系统实现中全部或部分采用智能化的处理技术或手段”均可称作智能语音处理(张雄伟 等, 2020: 5)。可见,语音处理领域也存在指称术语滞后于智能化的

现实,因而迫切需要一个能统合经典语音处理与智能语音处理两方面实践的概念术语。

再如,耿立波等(2014)提出了“机器语言能力研究”的概念。他们给出的解释是,机器语言能力主要指机器对人类自然语言信息的智能化处理能力,研究内容涉及机器学习、机器翻译、信息检索、人机问答、语言文字视听信息的机器自动化处理以及物联网中机器与机器、机器与人之间语言信号的传感等诸多方向。机器语言能力研究的核心是探索如何赋予机器以人的语言能力,使机器能够模仿人脑的语言加工机制,生成、理解和学习人类自然语言,实现机器与人、机器与机器之间的有效交际。显而易见,这一概念虽然明面上没有“智能”字样,但在内涵上差不多就是本文所称的“语言人工智能研究”的同义语。正如他们自己所说,“机器语言能力是一个融计算机科学、人工智能、自动化控制、数学、语言学、脑科学、认知科学等多门学科为一体的现代交叉科学研究领域”(耿立波 等, 2014: 34)。

最值得一提的是周建设先生提出的“语言智能”(Language Intelligence)概念(吕学强 等, 2017)。该概念于 2013 年提出,被界定为人工智能范畴的一个专门术语,其简明定义是:语言智能是语言信息的智能化,是运用计算机信息技术模仿人类的智能,分析和处理人类语言的过程,是人工智能的重要组成部分及人机交互认知的重要基础和手段(周建设 等, 2017)。后来他又深化、拓展了这一理解,认为语言智能是基于人脑生理结构和言语认知神经运作机理,利用大数据与人工智能技术,全面认识自然语言属性,对语言信息进行抽取、加工、存储和特征分析,同构人机意识关系模型,让机器模仿人类自然语言活动,实施类人言语行为,让机器具备听、说、读、写、译、评的能力,最终达到人机语言自由交互(姜孟 等, 2023;李西 等, 2021)。周建设(2023)还明确指出了语言智能的学科性质。他认为,语言智能就是机器理解并模仿人说话的科学,是研究人类语言与机器语言之间同构关系的科学,是一个融合神经科学、认知科学、思维科学、哲学、逻辑学、心理学、语言学、计算机科学等多个学科的新兴交叉学科。作为国家的一种新兴学科,语言智能研究集中在语言智能理论、语言智能技术和语言智能应用三个方向,旨在“发展语言智能科技,培养语言智能人才,推进语言学科教育智能化,促进教育高质量发展,助力国民语言能力提升和人文素养提升”(周建设, 2023: 2)。可见,语言智能在所指上也是与“计算语言学”“自然语言处理”“智能语言处理”“机器语言能力”等差不多的同义概念,但它独辟蹊径,从智能高度来看待人的语言能力,至少有两个优势:(1)“语言智能”天生包纳人类语言能力自然之维与人工之维,即“语言自然智能”与“语言人工智能”的双重含义,具有术语统摄优势。(2)这一概念有望超越前述“计算语言学”“自然语言处理”等概念所内蕴的历史方位的局限性,更好地匹配、相称于当今语言人工智能研究的广阔现实,具有学科统摄优势。

综上,近年来,语言人工智能研究在术语名称上表现出了不满现实、向“智能”聚考、向“智能科学”傍依的趋势与走向。这一现象所折射的是,语言人工智能的研究实践已经远远

走在了指称该学科领域的现有名称术语之前,“计算语言学”“自然语言处理”等名称术语已经滞后落伍,名实已不相符。“名不符实”必然要求“名实相符”。这表明,语言人工智能研究正在迎来一个崭新的历史方位——第五个历史方位。

3.2 从学科之实来看第五个历史方位

从学科之实来看,前四个历史方位代表着人类为攻克、实现人的离身语言智能之梦而发起的四次大尝试、大冲锋,成绩卓著。

第一个历史方位“思想乌托邦”(前机器翻译)时期产生了“引”人的翻译智能于人的肉身之外的奇思妙想,完成了逐梦中的思想先行、不怕做不到就怕想不到的历史跨越。第二个历史方位“泛机械”(机器翻译)时期以新诞生的计算机为机械装置的典型代表,把第一个历史方位中产生的奇思妙想变成了初步现实,完成了万事开头难、不仅想到还真正做到的第二步历史跨越。第三个历史方位“语言学主导的符号主义”(计算语言学)时期把“离身翻译智能”之梦扩展为“离身语言智能”之梦,并在挫折中尝试傍依语言学,以其为靠山,于无路可走中走符号主义之路,完成了从小到大、从无依无靠到有学科依靠的第三步历史跨越。第四个历史方位“计算机科学主导的连接主义”(自然语言处理)时期在傍依语言学遭遇困难后,尝试改寻计算机科学为靠山,以概率统计为主导,绕开语言的语义难题,走连接主义之路,完成了东方不亮西方亮、天无绝人之路的第四步历史跨越。然而,科学的攀登永无止境,在经历过四次大跨越、取得卓著成绩的同时,语言人工智能研究也来到了深水区,面临更深层、更具挑战性的问题与困难,其解决需要在新的历史方位寻觅新的靠山,尝试新的路径,实现新的跨越。

放眼当下,语言人工智能研究在语言智能的内在机制与底层原理上着力有限,所取得的突破性进展不多,整体上采取的是一条“绕道走”的研究路线。首先,从研究的内容实质与达到的目标高度来看,当前的语言人工智能研究还处于“有多少人工就有多少智能”阶段,距离真正的人的语言自然智能还有相当距离。今天,在语言人工智能研究领域里最具神通、最有影响的研究方法是基于统计的自然语言处理与基于人工神经网络的自然语言处理。但无论哪一种方法,它们所做的都还只是“浅层的”自然语言处理,相关技术还只能支持完成“浅层句法”或“简单标记”任务,对于更复杂的语言现象理解、语义关系抽取以及更专业的语言资料处理还一筹莫展。实际上,这与计算机本身的性质与原理有关。计算机是计算的机器,它以浮点数为输入和输出,擅长执行加、减、乘、除之类的计算。自然语言本身并不是浮点数,为了能够存储和显示自然语言,计算机需要把自然语言中的字符转换为一个固定长度(或者变长)的二进制编码。由于这个编码本身不是数字,对这个编码的计算往往不具备数学和物理含义。这意味着,一方面,如何让计算机有效地“计算”自然语言,是计算机科学家和工程师面临的永久难题和永无止境的追求;另一方面今天的计算机对自然语言的处理仅是对代表语言的二进制符号串的一种操作,并未理解语言本身的语义和涵义。此外,自然语言处理从浅

出走向深层,从不理解走向真正的理解,还涉及更复杂的情感计算、常识计算、知识图谱和不确切性信息处理等方面的技术(廉师友, 2020),这些都构成巨大的挑战。徐英瑾(2022)也认为,目前语言人工智能研究存在四大不足:(1)不同的自然语言处理机制之间缺乏融合;(2)自然语言处理技术与人工智能研究的其他技术缺乏彼此融合;(3)基于大数据的自然语言处理技术的运作必须以"剥削"人类的智能为前提;(4)基于大数据的自然语言处理技术缺乏灵活处理隐喻、反讽、双关等修辞现象的能力。言下之意,今天计算机所展示的语言智能在本质上只是人类智能的"反光映照体",就好比月亮只是太阳的"反光映照体"一样,距离语言智能的根基还很远。

再从研究的学科理路与构架体系来看,语言智能研究在四个历史方位中所经历的起起落落基本上只是应声于人工智能研究的起起落落(莫宏伟 等, 2020),尚未形成一个理论系统、方法成熟、技术完备的学科领域格局与面貌,独立自主性不够。综观整个自然语言处理领域,还远未建立起"一套完整、系统的理论框架体系",不少理论研究还只是处于盲目的探索阶段,对一些新的机器学习方法或未曾使用过的数学模型的尝试,还带有很大的主观性和盲目性(宗成庆,2019: 15-16)。在技术实现上,虽有许多改进,但这些改进要么限于"对一些边角问题的修修补补",要么只是针对特定条件下一些具体问题的处理,未能从根本上建立一套"广泛实用的、鲁棒的处理策略"(宗成庆, 2019: 15-16)。从研究现状来看,"自然语言理解和处理的理论体系尚未真正建立,技术方法仍然十分初步"(宗成庆, 2019: 3)。

钟义信(2018)在检视人工智能研究的不足时指出,现行人工智能研究路径或者是结构主义(模拟脑的结构,也即连接主义),或者是功能主义(模拟脑的功能,也即符号主义),或者是行为主义(模拟智能系统的行为,也即感知动作系统),尚未实现统一。他认为,目前不同范式的人工智能以不同的认识论学派为根基,所形成的是各种专用的人工智能。要实现通用的人工智能,需要融合各派的认识论作为理论基础,需要以符号主义、连接主义和行为主义不同哲学路向加以贯通整合的新型认识论为导引。要致力于创建"结构—功能—行为"整合融通、"意识—情感—理智"三位一体的通用人工智能理论。他把这种新的研究理论与范式称作"机制主义"。笔者认为,这也适合于当下的语言人工智能研究。针对当前的问题与瓶颈,语言人工智能研究需要主动走进新的历史方位,以当今的智能科学为靠山,立足语言作为人的智能的本质与机理,从其底层原理入手,系统建构语言人工智能独立的学科逻辑、理论体系与方法技术体系,走一条新的"机制主义"研究之路。

"自然语言处理毕竟是认知科学、语言学和计算机科学等多学科交叉的复杂问题,当我们从外层(或表层)研究语言理解的理论方法和数学模型的同时,不应该忽略从内层揭示人类理解语言机制的秘密,从人类认知机理和智能的本质上为自然语言处理寻求依据。"(宗成庆,2019: 16)以色列学者舒利·温特尔(Shuly Wintner)(2009)也批评了当今语言人

工智能研究对语言机制的严重忽视。她指出,当前的自然语言处理工程已经把语言学看得可有可无,研究的几乎都是程序技术或算法问题,很少关注自然语言处理工程背后隐藏的语言本身的基础性问题。20 年前,研究者们往往既对开发自然语言处理的应用系统感兴趣,也对语言学过程的形式化以及自动推理感兴趣。现今,他们却往往只对开发自然语言处理的应用系统感兴趣,对于语言学过程的形式化以及自动推理等方面的研究颇有些不屑了。为此,她呼吁应当让语言学重返计算语言学,不能让计算语言学成为没有语言学支持的计算语言学。冯志伟先生也呼吁,在自然语言处理研究中,单纯采取像蜘蛛那样的理性主义(符号主义)的方法,只依靠规则(只凭自己的材料来织成丝网),或单纯采取像蚂蚁那样的经验主义(连接主义)的方法,只依靠统计(只会采集和使用),都是不对的;而应当像蜜蜂那样,把理性主义和经验主义两种机能更加紧密地、更精纯地结合起来(既在庭院和田野里从花朵中采集材料,又用自己的能力加以变化和消化),推动自然语言处理的进步(宗成庆, 2019: 15-17)。“理性主义”与“经验主义”相结合的方法,实际上强调的就是要从内在机制、底层原理上下功夫,要迎难而上,不能完全靠概率统计的机巧、靠遇难绕道走来过日子。

“机制主义”研究之路的实质是从智能科学的宏大视野与高度来探索语言人工智能。语言具有复杂深邃的心脑机制,以其为观照,方能真正洞察当今语言人工智能研究的得失。从(心理)语言学的角度,人的语言能力一般分为语言习得、理解与产出能力。由是观之,当下对语言人工智能的研究主要聚焦于语言理解与语言产出,分别被称为“自然语言理解”(Natural Language Understanding, 简称:NLU)与“自然语言生成”(Natural Language Generation,简称:NLG)。自然语言理解是让计算机通过各种分析与处理理解人类的自然语言(包括其内在含义),而自然语言生成则关注如何让计算机自动生成人类可以理解的自然语言形式或系统(黄河燕 等, 2020: 2)。宗成庆(2019)列举了当今自然语言处理研究的 16 项内容,其中有 13 项内容对应于人的语言理解能力,仅有三项对应于人的语言产出能力,尚无内容对应于人的语言习得能力(见表 1)。这一方面反映了当下语言人工智能研究确实偏重于人的语言理解能力,对人的总体语言能力的模仿与再现的水平还不够高,另一方面也表明语言人工智能研究要实现根本性的突破,绕不开人类语言习得、理解与产出的复杂机理与机制。实际上,还可基于认知科学的立场,从人类语言运作的认知系统或概念系统的角度,将语言能力区分为概念能力、语词能力与言语外化能力。顺此思路,无论是语言理解能力、语言产出能力还是语言习得能力,都还可以析解到更下位的层面,即:概念子能力、语词子能力与言语外化子能力(姜孟 等, 2015; 姜孟, 2024)。这进一步表明,语言深层的心脑机制复杂广大,是统领语言人工智能各层面、各维度研究的总纲,纲举才能目张。

表 1 NLP 研究内容与人语言能力的对应一览

序号	自然语言处理的研究内容	所对应的人的语言能力
1	信息检索	语言理解
2	信息过滤	
3	信息抽取	
4	机器翻译	语言生成
5	自动文摘	
6	文档分类	语言理解
7	文本/数据挖掘	
8	舆情分析	
9	文字编辑与自动校对	
10	作文自动评分	
11	问答系统	语言生成
12	隐喻计算	语言理解
13	光读字符识别	
14	语音识别	
15	文语转换	语言生成
16	说话人识别/认证/验证	语言理解

莫宏伟等(2020)在描述人工智能的学科目标与任务时指出,人工智能研究智能的机制与规律,致力于构造智能机器,是研究如何使机器具有智能的科学。从这一上位学科的视角出发,语言人工智能要研究语言智能的机制与规律,致力于构造语言智能机器,其研究范畴不仅涉及对人脑语言认知机理、语言习得与生成能力的探索,还包括对语言知识表达方式及其与现实世界之间的关系,语言自身的结构、现象、运用规律和演变过程,大量存在的不确定性和未知语言现象以及不同语言之间的语义关系等各方面问题的研究(宗成庆,2019)。由此,语言人工智能工程天生具有跨学科、多学科交叉属性,其研究上的突破必须依赖认知科学、计算机科学、语言学、数学与逻辑学、心理学等多学科的研究成果。当前,能统合多学科,引领语言人工智能研究走向机制主义道路的正是“智能科学”学科。

涂序彦(2019)指出:“智能”是“信息”的精彩结晶,“智能化”是“信息化”发展的新动向、新阶段,“智能科学技术”是“信息科学技术”的辉煌篇章,“智能科学技术”(intelligent science & technology, 简称:IST)研究广义的智能问题,是关于广义智能的理论方法和应用技术的综合性科学技术领域,其研究对象包括:自然智能(人的智能和其他生物的智能)、人工智能(机器智能和智能机器)、集成智能(人的智能与机器智能构成的人机互补的集成智能)、协同智能(指“个体智能”相互协调共生的群体协同智能)以及分布智能(如广域信息网、分散大系统的分布式智能)”。虽然他对“智能科学”的定义中没有明确提到“语言智

能”问题,但其“研究广义的智能问题”的定义中必然涵盖了人的语言能力。

周昌乐(2021:1)给智能科学的称谓是“智能科学技术学科”,并将其研究的对象与目标界定为:“将人类智能(部分地)植入机器,使其更加聪明灵活地服务于人类社会”。他认为,智能科学的内涵涉及智能哲学、智能科学、智能技术等多个方面,人的语言是智能科学理论研究的核心内容之一,占其三分之一的天下。用他自己的话说,“智能科学的理论主要运用计算理论,围绕人类心智能力,开展神经计算建模、认知程序模拟和自然语言处理三个方面的研究内容”(周昌乐, 2021: 2)。他所绘制的智能科学理论架构图如下(图 1):

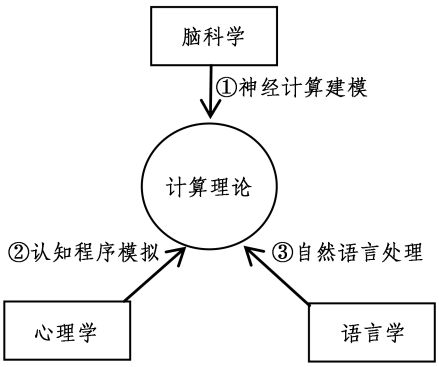

图 1　智能科学理论涉及的内容及其关系(周昌乐,2021:2)

史忠植(2019)专门著书对智能科学进行了构建与阐述。他认为,智能科学是由脑科学、认知科学、人工智能等创建的前沿交叉学科,旨在研究“智能的本质和实现技术”,是“生命科学的精华、信息科学技术的核心,现代科学技术的前沿和制高点,涉及自然科学的深层奥秘,触及哲学的基本命题”(史忠植, 2019: 3)。在智能科学的几个组成学科中,脑科学从分子水平、细胞水平、行为水平研究自然智能机理,建立脑模型,揭示人脑的本质;认知科学是研究人类感知、学习、记忆、思维、意识等人脑与心智活动过程的科学;人工智能研究用人工的方法和技术,模仿、延伸和扩展人的智能,实现机器智能。他认为,智能科学的目标是探索智能的本质,建立智能科学和新型智能系统的计算理论,解决对智能科学和信息科学具有重大意义的基础理论和智能系统实现的关键技术问题。它不仅要进行功能仿真,而且要从机理上研究和探索智能的新概念、新理论、新方法;不仅要运用推理,自顶向下,而且要通过学习,由底向上,两者并存。史忠植(2019)将语言智能视为智能科学研究的核心内容。根据他的界定,智能科学的研究内容包括:计算神经理论、认知计算,知识工程、自然语言处理、智能机器人等。此处,“自然语言处理”指的就是语言人工智能(史忠植, 2019: 10-13)。

概言之,从学科之“实”来看,语言人工智能研究已经走到了一个前所未有的深水区,到了直面人类语言智能的本质、内在机制与底层原理的时候了,“绕道走”的方法已经难以维系。与此同时,脑科学、生命科学等方面的新进展已经使智能科学应运而生。作为智能

科学的核心关切与主打领域方向之一,语言人工智能研究有望逐步走上“机制主义”的发展路线,并在智能科学的框架内成长为一个比较独立的分支学科——语言智能科学。这一切表明,语言人工智能研究已经走到了以智能科学为靠山的机制主义历史方位,可以简称为“语言智能”历史方位。

4 结语

语言是人的一种特别智能,致力于人类语言智能于人体之外机器实现的研究被称为语言人工智能研究。迄今,语言人工智能研究已经走过了 75 年,其间涌现了“机器翻译”“计算语言学”“自然语言处理”“自然语言理解”“语言智能”等多个相关名称术语。这些名称术语既各有所指又交叠瓜葛,令本领域的研究者和有志于进入本领域的研究者有些眼花缭乱、莫衷一是,更不利于洞察、把握本学科发展的历史逻辑与发展走势。

本文站在 10~30 年后的学科未来看历史与现今,透过学科指涉名称上的差异追寻语言人工智能 70 多年的风雨之路,解析其背后的思想理路、主线脉络、方法重心、技术逻辑、历时承继与未来趋向。在深入分析、考察与阐释有关史实与现实的基础上,提出新的主张与判断:语言人工智能研究已经走过了思想乌托邦(前机器翻译)、泛机械(机器翻译)、语言学主导的符号主义(计算语言学)、计算机科学主导的连接主义(自然语言处理)四个历史方位,正在迎来第五个崭新的历史方位,即智能科学主导的机制主义时期(见图 2)。学界新近提出的“语言智能”概念,是这一历史新方位恰当的代名词。

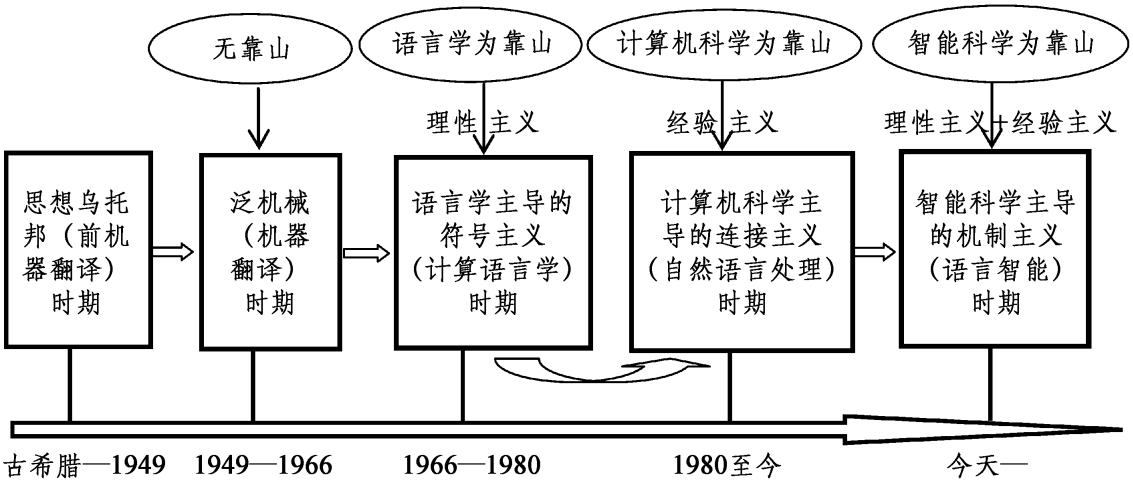

图 2 语言人工智能发展的五个历史方位

当前,语言智能学科领域仍处在不断流变、尚待定型、有容乃大的过程中。这一主张与判断的提出有助于撇开语言人工智能发展中的历史细节,抓大放小,异中求同,透过名称术语交叠混杂的表象看其背后的实质,厘清学科 70 多年行进的轨迹主线与思想理路,以更好地把握发展大势,能动地面向、塑造该学科的未来。

参考文献:

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural Machine[J]. Trans. by Jointly Learning to Align and Translate. *arXiv*:1409.0473v6 [cs.CL] (2015-04-24).

Bar-Hillel, Yehoshua. 1960. The Present Status of Automatic Translation of Languages[J]. *Advances in Computers* (1): 91-163.

Manaris, Bill. 1999. Natural Language Processing: A Human-Computer Interaction Perspective[J]. *Advances in Computers*, (47):1-66.

Shannon, C.E. 1948. A Mathematical Theory of Communication [J]. *Bell System Technical Journal* (27): 379-423.

Sutskever, Ilya, Oriol Vinyals & Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks[J]. *Advances in Neural Information Processing Systems* (4): 3104-3112.

Weaver, Warren. 1955. Translation[G]// William N. Locke & Andrew Donald Booth. *Machine Translation of Languages*: *Fourteen Essays*. Cambridge: MIT Press.

Wintner, S. 2009. What Science Underlies Natural Language Engineering? [J]. *Computational Linguistics* (4): 641-644.

毕然, 孙高峰, 周湘阳, 刘威威. 2022. 深度学习:零基础实践[M]. 北京: 清华大学出版社.

冯志伟. 1996. 自然语言的计算机处理[M]. 上海: 上海外语教育出版社.

冯志伟. 2011. 计算语言学的历史回顾与现状分析[J]. 外国语 (1): 9-17.

冯志伟. 2019. 我国计算语言学研究 70 年[J]. 语言教育 (4): 19-29.

耿立波, 刘涛, 俞士汶, 孙茂松, 杨亦鸣. 2014. 当代机器语言能力的研究现状与展望[J]. 语言科学(13): 34-41.

黄河燕, 史树敏, 贾珈, 黄民烈, 韩先培, 刘洋, 刘奕群. 2020. 人工智能:语言智能处理[M]. 北京: 电子工业出版社.

姜孟, 王霞, 潘雪瑶. 2023. 语言智能新文科建设与发展探索[G]//周建设. 语言智能研究. 天津:天津大学出版社, 7-13.

姜孟, 周清. 2015. 语言概念能力假设与外语学习者的“隐性不地道现象”[J]. 外语与外语教学 (4): 43-49.

姜孟. 2024. 第二语言概念能力探索:第二语言习得研究的概念进路[M]. 北京: 北京大学出版社.

李德毅. 2023. 机器具身交互智能[EB/OL]. 智能系统学报. (2023-02-03)[2024-03-10]. https://www.csgpc.org/detail/20068.html.

李德毅. 2024. 论智能的困扰和释放[J]. 智能系统学报 (1): 249-257.

李沐, 刘树杰, 张冬冬, 周明. 2019. 机器翻译[M]. 北京:高等教育出版社.

李西, 王霞, 姜孟. 2021. 语言智能,赋能未来:第五届中国语言智能大会综述[J]. 外国语文 (2): 141-144.

李宇明. 2023. 语言智能与社会进步[EB/OL]. 北京语言文字工作协会. (2023-01-19)[2024-03-12]. http://www.bjywxh.org.cn/html/2023/hydt_0119/1957.html

李珍. 2020. 人工智能的自然之维[J]. 云南社会科学 (1):40-46.

廉师友. 2020. 人工智能导论[M]. 北京:清华大学出版社.

刘颖. 2002. 计算语言学[M]. 北京: 清华大学出版社.

刘涌泉. 1963. 机器翻译和文字改革(上)[J]. 文字改革 (2):1-3.

沈家煊. 2000. 现代语言学词典[M]. 北京: 商务印书馆. 译自 Crystal, David. 1997. *A Dictionary of Linguistics and Phonetics* (4th ed.)[M]. Oxford: Blackwell.

史忠植. 2019. 智能科学(第三版)[M]. 北京: 清华大学出版社.

孙茂松, 周建设. 2016. 从机器翻译历程看自然语言处理研究的发展策略[J]. 语言战略研究 (6): 12-18.

涂序彦. 2019. 智能科学技术著作丛书序言[G]// 史忠植. 高级人工智能. 北京: 科学出版社.

徐英瑾. 2022. 人工智能如何“说人话”？——对于自然语言处理研究的哲学反思[J]. 自然辩证法通讯 (4):1.
亚里士多德. 1982. 物理学[M]. 张竹明,译. 北京:商务印书馆.
约翰·塞尔. 2006. 心、脑、科学[M]. 杨音莱,译. 上海:上海译文出版社.
张雄伟, 孙蒙, 杨吉斌. 2020. 智能语音处理[M]. 北京:机械工业出版社.
郑捷. 2019. NLP 汉语自然语言处理:原理与实践[M]. 北京:电子工业出版社.
钟义信. 2018. 机制主义人工智能理论——一种通用的人工智能理论[J]. 智能系统学报(1): 2-18.
周昌乐. 2021. 智能科学技术导论[M]. 北京: 机械工业出版社.
周建设, 吕学强, 史金生,等. 2017. 语言智能研究渐成热点[N]. 中国社会科学报,2017-02-07.
宗成庆. 2019. 统计自然语言处理[M]. 北京:清华大学出版社.

Language Intelligence as the Historical New Milestone of Artificial Language Intelligence Research: The Theoretical and Methodological Construction of Language Intelligence Science (IV)

JIANG Meng

Abstract: Language, as as a special human intelligence, is divided into Natural Language Intelligence and Artificial Language Intelligence. While the former is embodied in the flesh, the latter is realized in the mechanical devices and is thus disembodied from the flesh. Nowadays, while linguistics is dedicated specifically to addressing Natural Language Intelligence, a couple of disciplines are dedicated to addressing Artificial Language Intelligence, including Machine Translation, Computational Linguistics, and Natural Language Processing. The present study, by approaching this field of inquiry from the viewpoint of the discipline's growth in 10-30 years, examined the mainstream thoughts, theoretical timeline, methodological focus, technological logic, historical inheritance and future trends. A new tenet and claim was proposed that research on artificial language intelligence, which has gone through four historical milestones, namely Thought Utopia (Pre-Machine Translation), Pan-Machinerism (Machine Translation), Linguistics-dominated Symbolism (Computational Linguistics), Computer Science-dominated Connectionism (Natural Language Processing), is now embracing a brand new fifth milestone, namely, *Intelligence Science-dominated Mechanism*. In this respect, the newly-proposed concept *Language Intelligence* is the exact designator of this new historical peak.

Key words: natural language intelligence; artificial language intelligence; development peak; mechanism

责任编校:蒋勇军