

大模型研究之我观

文 / 宗成庆

本文首先简要回顾自 1947 年机器翻译 (machine translation, MT) 概念提出至今, 计算语言学 (computational linguistics, CL) 和自然语言处理 (natural language processing, NLP) 发展 70 多年的历程, 分析每个阶段主流技术方法的理论基础。然后阐述目前 CL/NLP 研究面临的问题和未来发展的方向, 最后分享作者对大模型研究的一点粗浅认识和思考。

1 NLP 发展历程

NLP 技术研究经过了 70 多年的发展历程, 可以大致将其划分为如图 1 所示的三个阶段。每个时期采用的理论方法各不相同。上个世纪 80 年代末期和 90 年代初期之前的阶段属于理性主义时期, 主要采用的是符号逻辑方法、语言学家的理论, 包括乔姆斯基 (N. Chomsky) 语法理论和其他语言学家提出的结构句法理论等都发挥了重要作用, 这种方法主要利用规则和知识库等分析和解决语言学问题, 因此也被称为基于规则的方法。从上个世纪 90 年初期至 2013、2014 年之前的大约 20 多年属于经验主义时期, 主要采用统计学习方法解决问题, 常用的方法和模型包括 n 元文法模型、支持向量机、隐马尔科夫模型、最大熵分类器和条件随机场等。2013 年年底端到端的机器翻译方法提出, 并得到广泛应用, 从此进入连接主义时期, 神经网络方法成为主流, 卷积神经网络、循环神经网络和长短时记忆模型、Transformer 等神经网络架构先后被提出和使用。不

管是经验主义方法还是连接主义方法, 都是以大规模数据为基础, 因此被称为数据驱动的方法。

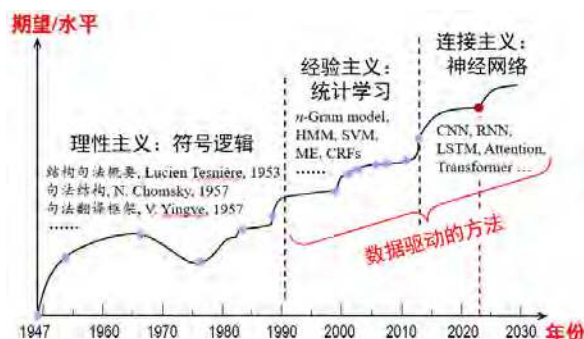


图 1 NLP 发展的三个阶段

最早的 NLP 研究是从机器翻译开始的。1946 年世界上第一台计算机研制成功之后, 美国洛克菲勒基金会自然科学部主任沃伦·韦弗 (Warren Weaver) 就提出了使用计算机开展语言翻译研究的设想, 但遭到了包括维纳在内的一些学界同仁的反对, 他们认为计算机只能做计算, 无法处理语言问题。但沃伦坚持自己的观点, 1949 年 7 月发表了以 *Translation* 为题的备忘录, 机器翻译概念被正式提出。1954 年, IBM 与美国乔治敦大学合作, 研发了第一个翻译展示系统, 实现了俄语和英语之间的简单翻译, 在随后的 10 年左右的时间里机器翻译研究成为研究热点。1956 年人工智能 (artificial intelligence, AI) 概念提出, 它比机器翻译概念的出现晚了将近 10 年。自然语言理解 (natural language understanding, NLU) 是人工智能研究的核心内容之一, 其目标是研究计算机如何像人一样理解自然

语言。

美国国家科学院在 1964 年成立了语言自动处理咨询委员会（ALPAC），对机器翻译研究进行了调研和评估，1966 年正式发布了调研报告，对机器翻译研究给出了令人失望的结论：机器翻译遇到了克服不了的障碍，即语义的表示和计算问题。如果语义无法表示和计算，怎么可能实现机器翻译？因此，报告认为对机器翻译研究给予更多的支持没有必要。该结论使世界范围内的机器翻译研究跌入冰谷。但是报告正式承认了计算语言学这个术语，得到美国国家科学院权威报告的认可，计算语言学便成为一门正式的学科方向走进了人们的视野。随着计算机网络的出现和移动通信技术的快速发展，自然语言处理概念应运而生。

自然语言理解、计算语言学和自然语言处理这三个术语的内涵和外延略有不同。自然语言理解更多地强调从语言智能的角度探索如何让计算机模拟人脑实现语言理解过程；计算语言学则更多地从数学建模的角度研究语言的处理问题；而自然语言处理则是从语言工程和应用系统研发的角度解决语言问题。在实际研发过程中我们很难严格地区分它们，也没有必要区分。因此，我们有时将其统称为人类语言技术（human language technology, HLT）。

传统的 NLP 技术研究通常从构成语言的两个基本要素入手，即从语音和文字两个角度开展相关技术的研究。从语音角度开展的研究包括语音识别、语音合成和说话人识别等，其主要处理对象是语音信号。我们通常将其称为语音技术。而以文字为主要处理对象的研究可以划分为两大部分，一部分主要涉及分析、理解技术，如文本的分类、情感分析、观点挖掘、信息检索、文本校对等；另一部分以归纳、转换和生成为主，如机器翻译、自动摘要、自动写作、问答系统等，当然，归纳和生成也离不开分析理解和处理。我们通常所说的 NLP 就是以语言文本为主要处理对象的研究。

在实际应用中有些任务既需要语音技术，也需要 NLP 技术。比如语音翻译，既涉及语音识别的问题，也存在语言理解和翻译的问题。另外还有一部分工作涉及图像和视频处理问题，如视频翻译，在一段视频中既有语音，也有视频（图像），有时候还有文字（字幕），为了将这段视频从一种语言翻译到另一种语言，既存在语音和文字识别理解问题，也存在视频理解、说话人肢体语言理解和情感分析等问题。再如，为了翻译一本纸质的书，需要首先利用 OCR 扫描技术识别书中的文字，然后再进行翻译，当然翻译过程中存在版面理解、图表理解和文字图像处理问题（当然，也可以不经过 OCR，直接实现端到端的文本图像翻译）。在这类任务中处理对象涉及语音、文本和图像三个模态信息，是一个地道的多模态信息处理问题。

除了上述从应用角度开展的 NLP 研究以外，还有一个基础性的问题，就是探索人脑是如何学习、记忆和理解语言的，人脑实现语言理解的神经基础和认知机理是什么？这需要结合神经科学和语言心理学等开展研究，我们通常把它叫做语言认知计算。我们通常所说的语言智能实际上涵盖了与语言相关的从科学问题探索到应用技术研发的各个方面。

2 问题思考

从上面的介绍可以看出，NLP 技术发展的每一个阶段采用的主流方法都不一样，从最早使用的模板方法，到后来的规则方法，经历了大约 40 年，而统计方法大约走过了 20 年。2014 年至今神经网络方法只有 10 年左右的时间。在 2017 年 Transformer 架构提出之前，NLP、语音技术、图像处理和计算机视觉等不同的研究方向都有各自不同的处理方法，甚至在 NLP 领域内机器翻译、文本情感分析和自动文摘等每一个具体方向的研究方法都不尽相同。但是，Transformer 被提出之后，几乎所有的研究方向都被统一在 Transformer 框架下。尤其自 2023 以来，

信息处理领域的所有研究任务、应用场景和模态信息几乎都采用大一统的处理模式——基于大语言模型的处理方法。

一年一度的国际计算语言学学术年会（简称“ACL 大会”）是 CL/NLP 领域顶级的国际学术会议。2025 年 ACL 大会于 7 月 27 日至 8 月 1 日在奥地利首都维也纳召开，根据大会程序委员会主席的统计，本次年会收到论文投稿 8 360 篇，在论文题目或摘要中含有“大语言模型 (LLMs)”及相关关键词（如 GPT、LlaMa、DeepSeek 等）的论文占投稿总数的 67%。本文作者大致估计，其余 33% 的论文中至少有一半与大语言模型相关。从会议现场展示的论文（包括口头报告和 Posters）情况看，非大（语言）模型相关的论文成为“稀缺”样本。也就是说，大约 85%~90% 甚至更多的论文工作在研究大模型或者使用大模型技术。除了 ACL 大会，CL/NLP 领域的其他会议，包括 EMNLP、NAACL、EACL、AAACL、COLING，甚至扩展到人工智能和机器学习领域的其他大会 AAAI、IJCAI、ICML、ICLR 等，都存在类似的情况。

纵观近年来 CL/NLP 领域的发展状况，我们不难得出如下结论。

（1）面向应用目标的技术研发已经从单一模态走向多模态信息处理。自大模型出现之后，进入 CL/NLP 领域几乎没有门槛，语音、图像和视频处理的研究门槛同样大幅度降低，多模态信息融合成为必然趋势。

（2）学术界与企业界的合作更加紧密。由于学术界在数据、算力和平台等方面的欠缺，要开展大模型研究不得不与企业界密切合作，随着数据量和模型规模的增加，学术界对企业界的依赖性越来越强。从某种意义上讲，学术界不得不在企业界的“绑架”和“裹挟”中艰难前行。

（3）相关领域的学术会议同质化现象愈加严重，大量雷同甚至重复性工作出现在同领域的相关会议

上。会议主题和组织形式的差异和特色越来越小。与此同时，面对大量涌现的稿件难以找到足够多的高水平审稿人，论文自动评审技术研发悄然兴起。可以大胆地预测，未来论文撰写和评审或将主要靠 AI 系统自动完成，作者只需要负责创新方法的设计、实验和稿件质量把关。

令人欣慰的是，中国学者近年来在 CL/NLP 领域获得了巨大的进步。20 年前中国学者在国际计算语言学和自然语言处理领域顶级学术会议上发表的论文数量还十分有限，在国际舞台的话语权微乎其微，而今天的状况已是天翻地覆。根据 2025 年 ACL 大会程序委员会主席的统计，51% 的投稿来自中国，最终获奖的杰出论文（outstanding papers）和最佳论文（best papers）数量中国学者占有半壁江山，数千人规模的 ACL2025 大会半数以上参会人员来自中国。这种巨大的进步所反映出来的一方面是中国学者奋发努力，勇追国际前沿不可阻挡的强大态势和成效；另一方面所体现的是我国经济实力和科研经费投入的巨大提升，这让我国学者拥有了更多的机会和信心。

但是，在追赶国际技术前沿和潮流的同时，我们必须冷静地面对现实，并认真地思考学科未来的发展和出路。大语言模型是 CL/NLP 任务的终极方法吗？换句话说，如果沿着这条路线走下去，CL/NLP 问题甚至通用人工智能问题就会被彻底解决吗？如果不能，那么下一步 CL/NLP 如何发展？当然，在任何一条技术路线走到终点之前，没有人能够给出答案，也不应该草率给出结论。但是，从学术研究的角度，探索不同的技术路线，突破现有框架提出创新思想，是技术发展和进步的基本规律。回想一下，如果当年在规则方法一统天下的时候，没有人从另外一个角度思考问题，如何会产生统计方法？同样，在统计方法风靡全球，神经网络遭受冷遇的时候，如果没有人坚持不懈地潜心研究神经网络，CL/NLP 如何会发展到今天？

从另一个角度思考，大模型是一种数据和算力牵引的技术，如果没有充足的数据、算力和财力，要想开展大模型研究几乎是不可能的，而这些资源恰恰不是高校和科研院所所具备的，这就导致了前面所述在某种意义上学术界被企业界“绑架”的尴尬局面，学术界除了为企业“打工”和输送人才，别无选择。那么，学术界的初衷、任务和使命是什么？如果只是在已有的框架下亦步亦趋，如何产出原创性成果？如何体现学术界存在的意义？

3 个人观点与相关研究

毋庸置疑，大模型技术的确是当前研究的前沿和热点，围绕大模型开展研究无可厚非，尤其企业界，面向实际应用开展大模型落地技术研究，具有重要意义，近年来越来越多的参与者也的确推动了大模型技术的快速发展和迭代。但是，如果整个学术界走同一条技术路线，这不是一种正常的学术生态。针对大模型存在的问题从不同的角度探索不同的理论方法和技术路线，建立更好的 NLP 方法，是学术界应有的做法。因此，我建议在给大模型研究给予支持的同时鼓励非大模型方法研究，让不同的研究思路 and 学术观点有生存的空间，给予机会甚至专门创造机会让其在学术舞台上自由探索。

我们可以思考一下，人脑是如何表征、记忆和理解语言的？有人做过统计，一个正在学习语言的儿童如果掌握了 100 个汉字，就能够表达 47% 的意思；如果他（她）掌握了 2 500 个汉字，就可以达到 97% 以上的意思；如果他（她）经过训练掌握了 4 000 个汉字，就能够表达和理解 99% 以上的意思，几乎没有语言交流的障碍。可是，现在的大模型训练需要多大规模的语料、多长的训练时间、耗费多少的计算资源？以最早发布的 ChatGPT 3.5 为例，它的训练语料规模达到 45 T，如果按照一本书 85 万字计算的话，45 T 大约相当于 2 800 万本图书，而大英博物馆的存书量才 1 400 多万册。也就是说，

ChatGPT 3.5 训练使用的语料规模相当于两个大英博物馆的存书量，即便在如此大规模的语料上训练出来的模型对语言的理解能力仍存在很多问题，所耗费的算力和经费且不说。

另外，人在理解语言时通常是通过视觉、听觉和触觉等多模态信息融合完成的，而且人一旦获得了语言理解和表达的能力，只要大脑不受损伤就会一直拥有这种能力，且能够随着生活阅历的丰富（可以认为接受更多的训练和测试数据），在与人和自然环境的交互过程中不断提升和进步。可以说人脑是一个轻量化、低能耗、绿色环保的智能体。

为此，近年来我关注的一个重要方向就是探索人脑理解语言的神经机制，构建受脑启发的绿色 CL/NLP 方法，或者借鉴对人脑认知机理的初步认识改进现有的大模型结构。

在脑语言认知机理探索方面，我们团队主要聚焦在语言的神经编码和解码两个方面。为了开展这项研究，尤其针对汉语的神经编解码研究，我们先后构建了五个大规模数据集，包括汉语词汇级的神经影像数据集（<https://doi.org/10.1038/s41597-022-01840-2>）、汉语连续故事的神经影像数据集（<https://doi.org/10.1038/s41597-022-01708-5>）、大规模汉语语义主观评分数据集（<https://doi.org/10.1038/s41597-023-01995-6>）、多模态认知基准数据集（<https://doi.org/10.1007/s10579-025-09843-2>）和用于多模态机器翻译、情感计算的多模态信息处理数据集（<https://aclanthology.org/2023.findings-emnlp.673.pdf>）。在此基础上我们进行了一系列理论和技术方法探索，先后提出了基于 fMRI 数据构建语义表征功能驱动的大脑皮层网络分区方法（<http://eyemind.psych.ac.cn/PublishedPaper/Zhang2025.pdf>）；词向量语义维度层次化分解方法（发表于自然语言处理实验方法国际会议 EMNLP 2025）和基于脑活动数据的语句生成方法（<https://aclanthology.org/2024.naacl-long.211.pdf>）等。

神经科学的研究发现，人脑的不同脑区存在着功能上的划分，即功能分化现象，例如位于枕叶、后顶叶和下颞叶附近的视觉皮层区域负责的视觉功能，中央前回附近的脑区负责躯体的感觉运动等。在语言认知科学领域，研究人员也发现不同语言的表征之间存在着一定的分离。受这种现象的启发，我们提出了一种针对 Transformer 多头自注意力机制功能分化的方法，研究 Transformer 中是否存在多任务训练下演化出类似的功能分化现象，以及如何利用这一现象提高基于 Transformer 预训练模型的能力（<https://aclanthology.org/2023.findings-emnlp.673.pdf>）。进一步，我们提出了语言功能分化的混合专家模型结构，以解决“多语言诅咒”问题。所谓的“多语言诅咒”是指在大模型训练过程中语言种类达到一定数量时模型的性能由缓慢上升转为下降，这与希望同一个模型能够处理语言的种类越多越好的目标背道而驰，似乎模型性能受到了语言数量的诅咒。为了解决这一问题，我们将语言分化策略引入到混合专家结构层，在扩大模型参数量的同时减少不同语言之间的竞争关系，以促进相似语言之间的正向信息迁移。在 128 种语言上进行的实验表明，我们方法可有效阻止“多语言诅咒”现象的发生（<https://arxiv.org/pdf/2507.08309>）。另外，语言学家发现，人在学习一种新的语言时来自母语的先验知识有助于帮助学习者理解新的语言。借鉴

这种发现，我们提出了将资源丰富语言的生成能力向低资源语言迁移的方法，使模型在跨语言推理和逻辑推理任务上的性能得到显著提升（<https://aclanthology.org/2024.naacl-long.211.pdf>）。当然，这些研究只是对 Transformer 架构的改进，或者只是大模型性能优化的尝试，并没有提出任何全新的理论方法和技术框架，但是我们的研究方向是明确的。

4 结束语

无论如何，计算语言学和自然语言处理研究的最终目标是构建真正具有自然语言理解能力的低能耗绿色技术方法，只要这个目标没有达到，就始终有研究的必要。当然，这是一个理想的或许遥远的目标。需要说明的是，我不反对研究大模型，但我不认为整个领域都要走同一条路；我也切身感受到了快速发展的大模型技术所带给我们的紧迫感，但我不认为大模型方法是 CL/NLP 的唯一方法和最佳方法。探索受脑启发的 CL/NLP 新方法也并非对大模型技术的彻底否定和摒弃，而是面对大模型存在的问题，从不同的角度、不同的技术路线寻求更优的方法。我始终认为，立足现实量力而行，开展前瞻性研究，引领技术发展，是学术界不变的使命。我们现在缺少的不是论文数量，而是勇于突破和敢于独辟蹊径的原创性思维。

（参考文献略）



宗成庆

欧洲科学院外籍院士，中国科学院自动化研究所研究员，IEEE/ACL/AAIA/CAAI/CCF Fellow。主要从事自然语言处理和语言认知计算等研究，主持国家自然科学基金重点项目、国家重点研发计划重点专项等 10 余项，出版专著 6 部、译著 2 部。2025 年担任本领域最具权威的学术组织——国际计算语言学学会主席 (ACL President)，同时担任中国中文信息处理学会副理事长。曾任亚洲自然语言处理学会 (AFNLP) 主席、顶级学术会议 ACL 2015 程序委员会主席和 ACL 2021 大会主席。荣获国家科技进步奖二等奖、北京市科技进步奖一等奖和教育部自然科学奖二等奖等，被评为北京市优秀教师、中国科学院优秀导师和中国科学院大学李佩教学名师。享受国务院政府特殊津贴。