

语言学与人工智能：从符号到意义的计算建模

詹卫东

“没有一门科学比语言学同人类的本质更接近了，因为它研究的对象是我们借以表达我们个性的手段，也是我们的思想得以形成，被人接受，并获得连续性的手段。”

“没有哪一个科学领域像语言学那样存在着如此之多的谬误，不仅存在着，而且还继续被当作真理传授着。”

这是美国哈佛大学语言学家德怀特·鲍林杰（Dwight Bolinger, 1970-1992，又译作鲍林格）在他那本广受好评的语言学科普著作《语言要略》第一版（1968 年）序言中写下的两句话。前一句强调了语言学的重要性；后一句又表达了对语言学可谓刻薄的负面评价。这对一般人来说很可能并不好理解。语言就好像空气和水一样是人类必需的，但同时又是再普通不过的东西，不就是平淡无奇的“说话”吗？为什么把语言学抬高到如此重要的位置？跟文学、历史、哲学等学科相比，语言学真的比这些学科都更接近人的本质？而如此重要的语言学，为什么在一个语言学家眼里，又会充斥着谬误呢？

要理解鲍林杰的话，其实根本上来讲，需要面对的是一个苏格拉底式的拷问，一个貌似简单但实际上却很难回答的问题——**什么是语言？**

在人工智能技术飞速发展的今天，能说会道的机器已经不是什么新鲜事了。既然机器都会说话了，那语言还能算是人的本质属性吗？

一 语言：有限手段的无限应用

把“语言”作为一个科学研究的对象，其实是很晚近的事情。美国斯坦福大学计算机科学家，人工智能的早期研究者之一特里·维诺格拉德（Terry Winograd）在《语言作为认知过程》（1983）一书中，借鉴托马斯·库恩（Thomas S. Kuhn, 1922-1996）的科学范式革命理论，把语言学的历史概括为四个范式：规范语言学（prescriptive linguistics），比较语言学（comparative linguistics），结构语言学（structural linguistics），生成语言学（generative linguistics）。其中第一个范式严格来说并不适合称为“语言学”，而更应称之为“语文学”（language arts 或 philology）。这个阶段从古代一直延续到 17、18 世纪，人类在这漫长的时间里对语言的认识可以说并没有真正触碰到“什么是语言”这一问题。直至 18 世纪比较语言学问世，19 世纪末 20 世纪初结构语言学登场，20 世纪中叶乔姆斯基生成语言学革命，“什么是语言”这个极为深刻的问题，才逐渐成为现代语言科学反复叩问的核心研究目标。

语言以文字形式记录下来，散播四方，传之后世，就会因为地易时移而发生语义解读的困难。古人关注语言，主要目的就是解决书面文本的释义问题，特别是历史、宗教、哲学经典中的文本释义。比如《春秋公羊传》对“二月，公侵宋。”（《春秋·庄公十年》）的释义：“曷为或言侵？或言伐？桷者曰侵，精者曰伐。战不言伐，围不言战，入不言围，灭不言入，书其重者也。”这段释义是说：“侵”就是粗略出兵象征性地打仗，“伐”则是认真准备实质性地进攻，比“伐”更严重的，叫“战”，比“战”更严重的，叫“围”，比“围”更严重的，叫“入”，比“入”更严重的，叫“灭”。《春秋》的习惯是选择最严重的那个词来描述战事（即“书其重者也”）。《公羊传》通过展示“侵”的相关词语的意义差异，让读者理解整句话的意思是：二月份，鲁庄公侵犯宋国，不过只是做了做样子，并没打什么大仗。古典语文学的研究是为经典文化传承服务的，其研究关注的核心问题，一言以蔽之就是：**这句话是什么意思？**即始终是对个别的、具体的字词句的意义阐释，并不重视把语言作为一个整体对象来加以系统研究。

随着 14、15 世纪欧洲文艺复兴、宗教改革等一系列重大事件的发生，世界历史开始进入一个崭新的时期。大航海时代、殖民扩张、古登堡活字印刷术推动了近代科学的迅速发展，拉开人类文明现代化的序幕。在这样波澜壮阔的历史背景下，人类对语言的认识也同时发生着深刻的变化。语言学者们不再满足于只是解释个别字句的意思，语言研究的视野在空间的广度、时间的长度以及结构的深度上发生了翻天覆地的大变化。于是，就有了维诺格拉德所总结的科学范式意义上的三次语言学大转型，每次转型，都可以用语言学家提出的关于语言的一个全新的问题作为代表。

比较语言学的问题可以概括为：**语言如何演化**？随着传教活动、殖民地跨文化交流和经贸往来的日益频繁，世界上原本彼此隔绝的语言之间开始了大规模的近距离接触。出于沟通需要，语言之间的对比成了语言学者的要务。欧洲人的语言研究开始从文本释义，跃升到宏观审视不同时空语言之间的关系。当时的情形恐怕用“不比不知道，一比吓一跳”来形容也不为过。1786 年，时任英属印度孟加拉最高法院法官的威廉·琼斯（William Jones, 1746 - 1794）在亚洲学会（Asiatic Society）发表的年会演讲中称：“梵语……比希腊语更完美，比拉丁语更丰富，……但是在动词词根和语法形式方面与这两种语言都有显著相似性，这不可能是偶然的，这种相似性如此显著，没有一个考察这三种语言的学者会不相信它们同出一源……”。这段话极具代表性，反映了比较语言学追溯语言来源的研究兴趣所在。科学思想、实证分析、系统观念开始伴随着比较语言学的具体研究成果，逐渐成为语言学者的共识。为了厘清语言间的谱系关系并构拟原始祖语，除了历史文献证据，对各种鲜活的语言运用状况的记录和描写，即第一手实证材料的掌握，也逐渐成为语言学者研究语言的首要任务。在这个过程中，越来越多的语言学者对语言共时（synchronic）系统的兴趣开始超过对语言的历时（diachronic）演变研究，为结构语言学的登场做足了铺垫。

结构语言学的问题可以概括为：**意思从何而来**？跟古典语文学着眼于个别字词句的释义不同，也跟比较语言学过于关注语言演化的原因和过程不同，结构主义语言学者更强调研究语言的当下和语言系统本身。结构主义语言学认为，只有建立起一套科学的分析方法，把语言看作是一个符号系统，根据系统的组合规则，通过对比离析出基本单位，小的单位组合成大的单位，由简单到复杂，最终才能形成完整的解释语句意义的可靠程序。比如要分析“阿 Q 没理由怕吴妈”跟“阿 Q 没理由地怕吴妈”两句的意义差异，就要系统地考察“动词性词组+地”结构的分布环境和表义功能，涵盖更多类似的结构（如“没办法+地”“没骨气+地”“没良心+地”“没心情+地”等等），所提炼的结构关系规则是系统层面的规律，而不是停留在个例层面的就事论事的释义。瑞士语言学家费迪南德·索绪尔（Ferdinand de Saussure, 1857-1913）首次提出“语言”（langue）和“言语”（parole）的区分，主张以前者替代后者作为现代语言学的理论研究对象，把语言符号跟其意义的联系，置于系统观念之下来审视，尝试用更稳定、更具规律性的“语言”系统，来驾驭表面上千变万化捉摸不定的个人的“言语”实践。

生成语言学的问题可以概括为：**句子如何生成**？诺姆·乔姆斯基（Noam Chomsky）于 20 世纪中叶提出的这个震撼性的问题，彻底改变了现代语言科学的面貌。乔姆斯基在索绪尔关于“语言”和“言语”区分的基础上进一步提出“语言能力”（competence）和“言语表现”（performance）的对立。这一对立跟索绪尔理论的最大不同在于，“语言能力”的主体是个体，而不是社会群体。语言研究应关注如何为人脑心智构建语言结构系统，该系统要能够解释人的无限句子生成能力，即人为什么能说出以前从未说过的话，为什么能听懂以前从未听过的话。下面是乔姆斯基举过的两个经典例句：

- (1) Colorless green ideas sleep furiously. 无色的绿色思想狂怒地睡觉。
- (2) Furiously sleep ideas green colorless. 狂怒地睡觉思想绿色无色的。

句子（1）和（2）都让人觉得“不知所云”，但英语母语者仍能分辨（1）合语法，（2）不合语法。因为句子（1）可以用正常的句调念出来，而句子（2）只能一字一顿地念。之所以

如此，是抽象的句法结构在起作用：句（1）中是有词组单位的，比如 green ideas, sleep furiously 都是合乎英语句法结构要求的词组，只是存在语义不当的问题。而句子（2）中每个单词都是独立存在的，跟其他单词不发生组合关系。句（1）容易记住，句（2）不容易记。自乔姆斯基之后，语言学研究材料中就开始越来越多地出现这类不正常的“负例”，而不再只是文献中的“正例”。因为以往的语言学关注的是已经说出来的句子——语言的外化成品，而乔姆斯基心目中的语言学研究对象，是产生句子的原因——人脑中的语言机制。这就好比“授人以鱼不如授人以渔”，生成语言学关心的是“渔”——人的语言生成能力。

乔姆斯基将人的语言能力高度概括为“有限手段的无限运用”（infinite use of finite means）。这一语言观，可以说是 17 世纪法国唯理主义哲学家勒内·笛卡尔（René Descartes, 1596-1650）和 19 世纪德国理性主义哲学家威廉·冯·洪堡特（Wilhelm von Humboldt, 1767-1835）关于语言的创造性与人类心智关系的思想在 20 世纪的升级版。洪堡特提出**语言是“活动”（energeia）而非“产品”（ergon），语言通过有限的形式（如语法、词汇）服务于无限的思维需求，这是人类精神的根本特征**。乔姆斯基站在巨人的肩膀上，引入数学的形式化方法，将哲学家的思想落实为语言学的操作手段，使语言学的研究重心从如何**理解**句子扩展为如何**生成**句子，首次把语言研究跟人脑心智结构联系起来，主张人的语言能力源于人脑的先天机制，语言学应超越外在的言语现象的描写，深入到心智之内，解释语言的无限生成性。

二 语言学：构建形式与意义对应的模型

有人说生成语言学的尽头是生物学，意指要从大脑生物基础层面来解释人的语言能力。不过目前生成语言学的发展还远未达到解释人脑如何生成句子的层次，而是像数学建模那样，用形式化方法建立基本范畴和规则，从而区分合法表达式与不合法表达式，并进而为合法表达式构建相应的语义解析式，在句子形式和意义之间建立对应关系。下面通过一个简单例子，展示生成式规则（production rule）如何以有限手段生成无限句子。

表 1：语句及其对应规则示例

序号	实例	生成式规则
①	两部作品	np → qp np
②	文学作品	np → np np
③	通俗作品	np → ap np
④	鲁迅的两部通俗文学作品	np → np de np

上表中的生成式规则①，意思是箭头左边的 np 由箭头右边的 qp 加 np 组合得到，即“两部作品”这个名词性成分 (np) 由“两部”这个数量成分 (qp) 加“作品”这个名词性成分 (np) 组合而成。运用规则②和③，得到“通俗文学作品”（“通俗”是 ap，“文学作品”是 np）；再运用规则①，得到“两部通俗文学作品”；再运用规则④，得到“鲁迅的两部通俗文学作品”。

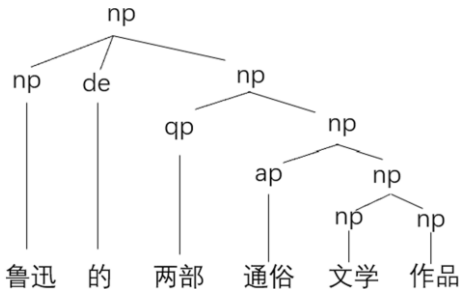


图 1：“鲁迅的两部通俗文学作品”的树结构示例

把运用生成式规则的过程记录下来，就是树结构。图 1 展示了表 1 例④的树结构图。从

上到下，最顶端的 np（树根）对应规则④，第二层最右端的 np 对应规则①，第三层最右端的 np 对应规则③，第四层最右端的 np 对应规则②。如果继续嵌套运用表 1 中的规则，还可以扩展出更长的组合：鲁迅的两部优秀通俗文学作品，鲁迅的两部优秀通俗文学作品的三种版本，鲁迅的两部优秀通俗文学作品的三种珍贵版本……。生成式规则的反复应用，使得语言可以由简到繁，从简单的实体开始，表达越来越复杂的思想。

生成式规则是有限的，而它们能生成的树结构理论上是无限的。生成语言学假设的这个模型，在落实过程中要面临的挑战很多，最主要的是：到底需要设置多少像 np，vp 这样的基础范畴，需要多少条规则，才能覆盖无限多的自然语言句子？更进一步，生成式规则不仅要能生成语法上的正例，还要避免生成语法上的负例。对于语法上的正例，语言模型还要判断它是否有合理的语义解释，以及是否有多个语义解释，即歧义（ambiguity）。生成式规则比较适合表达语法组合，至于语义解释，还需要借助语义规则（如谓词逻辑表达式，语义框架图模型等）来描述。限于篇幅，这里不展开讨论。下面通过一些实例来说明自然语言形式和意义对应关系的复杂性，以管窥豹，可以体会为自然语言建模的挑战有多大。

（一）同形异义现象

表 2：三段论推理示例

三段论推理示例（甲）	三段论推理示例（乙）
a) 人 是 会死的	a) 鲁迅的作品 是 不可能在一天之内读完的
b) 苏格拉底 是 人	b) 《祝福》 是 鲁迅的作品
c) 苏格拉底 是 会死的	c) 《祝福》 是 不可能在一天之内读完的

表 2 中甲例三段论推理是正确的。乙例形式上跟甲例一样，大前提和小前提单独看都对，但结论（c）显然是错的。问题出在哪儿呢？对乙例稍加分析后不难发现，“鲁迅的作品”在 a 中指“鲁迅的全部作品”，在 b 中指“鲁迅的一部作品”。这样，大小前提之间的推理联系就被切断了，导致三段论推理实际上无法成立。这个例子揭示了汉语中普通名词性词组存在量化歧义（quantificational ambiguity）的现象。显然，如果要描述“鲁迅的作品”的量化歧义，表 1 的产生式规则④就需要加以改进。

下面是跟“鲁迅的作品”相关的一些正负例和歧义例子，可进一步说明对规则的挑战。

- (1) a 两篇鲁迅的作品 b 鲁迅两篇的作品 c 鲁迅的两篇作品
(2) a 美术鲁迅的作品 b 鲁迅美术的作品 c 鲁迅的美术作品
(3) a 讽刺鲁迅的作品 b 鲁迅讽刺的作品 c 鲁迅的讽刺作品

三组例子中例 1b、2a、2b 不合语法，其余都是合法词组。例 1 和例 2 都可以由表 1 的规则生成。但要避免生成不合法的负例，同样也要改进规则。此外，表 1 现有规则无法生成例 3，需要补充新的规则：vp → vp np，np → vp de np。vp 代表动词性词组。3a 有两种语义，对应如下两个树结构。换言之，歧义可以用不同的树结构加以区分：

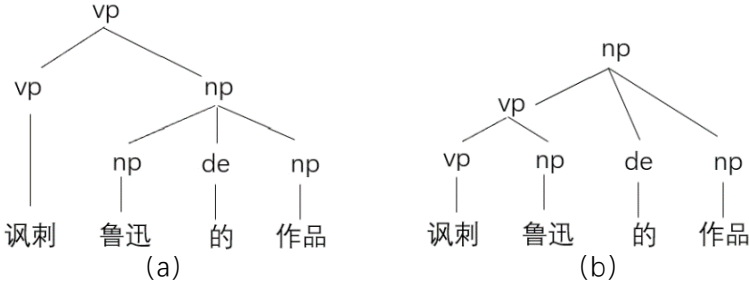


图 2：“讽刺鲁迅的作品”的两种句法结构树

“鲁迅的作品”在图 2a 中是否有量化歧义？能否也用树结构区分歧义？请读者思考。

（二）异形同义现象

(4) a 阿 Q 看不起未庄人

b 未庄人看不起阿 Q

(5) a 天空洒满阳光

b 阳光洒满天空

例 4a 和 4b 中的“阿 Q”和“未庄人”是逆序换位关系。整句意思不同。类似“狗咬人 ≠ 人咬狗”，代表了语言表达的常态。例 5a 和 5b 的形式差异，跟例 4 一样，也是谓词前后的主宾换位，但是，5a 和 5b 基本为同义关系。像例 5 这样的语言现象，汉语中还有很多。如：客厅挂这幅油画—这幅油画挂客厅，鸡蛋炒西红柿—西红柿炒鸡蛋，病毒感染人体—人体感染病毒，阿 Q 挨着吴妈—吴妈挨着阿 Q，杭州是浙江的省会—浙江的省会是杭州，一辆车挤六个人—六个人挤一辆车，我想死你了一—你想死我了……

显然，语言中既存在像例 4 这样主宾不可换位（异形异义）的常态情况，也存在像例 5 这样主宾可换位（异形同义）的超常情况。那么，主宾可换位句有多少种不同的类型？跟“洒满”同类的，可造成主宾换位异形同义句的动词有多少？为什么会有主宾可换位现象？本文无法展开讨论这些问题。下面仅通过例 6-8 简要展示问题的复杂性。

(6) a 香港澳门回归祖国

b 祖国回归香港澳门

(7) I want to call for **the return of linguistics to computational linguistics**.

a 我想呼吁**语言学回归计算语言学**

b 我想呼吁**计算语言学回归语言学**

(8) a 让舞台回归传统

b 让传统回归舞台

就例 6 而言，“回归”跟“看不起”同类，其构成的句子主宾不可换位，6a 成立，6b 不成立。但是有人在翻译例 7 英文句子时，直译为 7a，却引发了疑问，质疑者觉得译为 7b 更恰当。因为在计算语言学发展过程中，存在越来越不关注语言学研究的现象，因此要呼吁学界重视这个问题。在这样的背景下，译文应该是“计算语言学回归语言学”，而不是“语言学回归计算语言学”。事实上，译文 7a 和 7b 的语义区别没有 6a 和 6b 的区别那么大。对此，例 8a 和 8b 的对比或许可以佐证。例 8 中“传统”更侧重属性特征，“舞台”指一个空间场所，二者的语义关系是表达某种实体（及其典型特征）融入到某个空间（类似例 5）。例 7 中“语言学”也可以作为属性特征来理解（相当于“语言研究”），这样，“语言学”和“计算语言学”的关系，就不再是两个学术领域之间的关系，而是属性特征跟空间之间的关系。7a 和 7b 因此也可以看作是异形同义关系。

以上两类例子所展示的自然语言的复杂性，或可说明，语言学家试图构建能够生成（实际上也包含了理解）自然语句的理论模型，难度有多大。而从不同角度尝试构建这样的模型，也很容易让人联想到盲人摸象的隐喻，导致语言学中的不同理论体系，在观点互相对立的学者眼中，就会产生如鲍林杰教授所说的：语言学中存在着如此之多的谬误。或许，把“谬误”换成“争议”，更公允一些。如何让语言学从争议的泥潭中解脱出来，是语言学人都应关心的核心议题。要解决争议，就要有验证模型好坏的客观标准。人工智能的研究，特别是研究如何让计算机具备人类语言能力，既为实用，同时也是检验语言学理论的可行途径之一。

三 人工智能：操作语符的“模仿游戏”

计算机科学先驱英国数学家阿兰·图灵 (Alan Turing, 1912-1954) 在《计算机器与智能》(1950) 一文中提议用“模仿游戏” (imitation game) 来替代“机器会思考吗？”这个不太容易清晰定义的问题。图灵所说的“模仿”，具体就是让机器模仿人的语言活动，且能达到跟人自

如交流的水平，以至于在“隔墙笔谈”的假想场景下，对谈者无法区分交流对象是人还是机器（后人把这个游戏改称为“图灵测试”）。图灵在文末写到：“我们可以期待，总有一天机器会同人在所有的智能领域里竞争起来。但是，如何开始呢？这是一个很难决定的问题。许多人以为最好从下棋之类的极为抽象的活动开始，不过，还有一种办法也应加以考虑，就是为机器装配上容易买到的传感器，然后教机器理解并且说英语。这可以仿效通常小孩子学话的方式进行，比如以手指物称名等。我不能确定到底哪个出发点更好，但应该都值得一试。”

从发明初衷以及顾名思义来说，计算机似应只用于“计算”，computer 在图灵的 1950 年代还是指从事计算工作的人（即“计算员”），所以图灵文章的标题用 computing machinery 来指计算机。不过，超越时代的图灵告诉人们，“计算”并不限于数值计算，而是可以拓展到任意符号（包括文字、声音、图像）的信息处理。万物皆数。计算机中所有的问题最终都可以归结为有限个 0、1 分类问题的组合。机械计算有能力模仿“思考”。图灵在遐想计算机“思考”的实例时，既包括计算“34957 加 70764 等于多少”，也包括“以福斯桥为主题写一首十四行诗”。几乎自 20 世纪中叶电子计算机刚问世起，人类就开始了让计算机模仿人类语言活动的不懈探索。

计算机模仿人类语言活动的主要方法可以分为两大类，一是知识驱动，一是数据驱动。2024 年诺贝尔物理学奖得主杰弗里·辛顿 (Geoffrey E. Hinton) 站在“智能研究范式”的高度，对这两类方法做了精辟的阐释：前者是受逻辑学启发的思路，这一派人认为智能的本质是推理，而推理是可以符号规则操纵符号表达式来实现的。开发智能要做的，就是了解知识是如何表征的，是用什么类型的逻辑编程语言来表征的。学习之类的事可以等到以后再做；后者是受生物学启发的思路，这一派人认为智能的本质是学习神经网络中连接的强度。逻辑和其他事情要晚得多。得先了解人是如何学会控制自己的身体或识别物体之类的事情。推理之类的事以后再说。

这两种方法在“人如何教计算机”的具体做法上大不相同。

知识驱动派是依靠人力去提炼语言学知识，并以形式化的方法来把人类知识（如上文的生成式规则）整理为知识库、数据库、标注语料库等等教给计算机，这些知识既是人类易懂的，也是计算机可读的。计算机基于人类知识来模拟人的语言活动（如机器翻译、信息抽取、文本摘要等等），类似于围棋程序遵循人类棋谱来下棋。

数据驱动派是依靠计算机算力，设计恰当的人工神经网络架构，由计算机在海量的语言数据（目前规模达万亿字节量级）中观察文本符号的分布模式，训练神经网络模型参数（目前规模达千亿数量级），从而能够在给定上文条件下，预测最优的下文，实现模拟人类语言行为的效果，类似于围棋程序不依靠人类棋谱，而是通过自我对弈来训练其神经网络参数获得下棋能力。

下面通过两个小例子，对知识驱动和数据驱动两条路线各自的特点再略作阐释。

第一个例子是维诺格拉德在其博士论文（1971）中举的例子，目的是说明，理解语言除语言结构本身的知识，还需要用到语言外的知识，比如人类经验和常识。2011 年，加拿大多伦多大学的研究人员借鉴此例的思想，构建了名为“Winograd Schema Challenge (WSC)”的指代消解任务，用于测试计算机的常识理解能力，任务形式类似于问计算机下面两句中 they 指句中的哪个成分。如果是英汉机器翻译任务，则 they 在两句中分别要译为“他们”和“她们”。

- (1) **The city councilmen** refused to give the women a permit for a demonstration because **they** feared violence.
- (2) The city councilmen refused to give **the women** a permit for a demonstration because **they** advocated revolution.

答案是：they 在例 1 中回指“市议员”；在例 2 中回指“那些女人”。知识驱动派的解题思

路可以概括为：处于相同句法位置的代词“they”回指上文不同的成分，其背后是遵循同样的语义原则：主句中哪个成分能跟从句谓语语义相配，哪个成分就是从从句主语“they”替代的对象。根据常理，主句的两个指人名词性成分，跟从句谓语的搭配选择是不同的。例 1 中“害怕暴力”的是裁量游行活动能否举办的“市议员”。例 2 中“鼓吹革命”的是申请游行活动的“那些女人”。因此，根据回指的语义相谐原则，“they”在例 1 中就回指“害怕暴力”的“市议员”，在例 2 中则回指“鼓吹革命”的“那些女人”。数据驱动派的办法则是：将 they 的词向量（word vector）跟句中名词的词向量进行向量空间中的语义距离计算，例 1 中 they 的词向量因显著受到紧邻的“害怕暴力”影响，跟“市议员”的语义距离会更近；例 2 中 they 的词向量则显著受到紧邻的“鼓吹革命”影响，跟“那些女人”的语义距离会更近。语义距离更近的名词性成分，就是 they 指代的成分。

第二个例子是考察计算机对语言结构综合理解能力的一个小测试题：

“讽刺鲁迅的作品”是汉语的一个词组。现在替换“讽刺”这个词，生成新的词组，即在“（ ）鲁迅的作品”括号位置填入新的词语。要求：“讽刺鲁迅的作品”有两个意思，这两个意思的词组内部结构不同。替换后新词组的结构要保持跟原词组相同，也要有两个意思。请问，下面哪些词语满足要求，哪些不满足？

- (1) 收藏 (2) 拍卖 (3) 批 (4) 评 (5) 关于
(6) 赞美 (7) 青年 (8) 创意 (9) 创新 (10) 创举

答案是：(3) (4) (5) (6)。其他词语填入括号位置形成的词组，要么只有一个结构，一个意思解读（如 1、2、7），要么不能成立，是非法组合（如 8、9、10）。

对于这样的语言理解类任务，知识驱动派的做法是由人来挖掘语言形式和意义对应关系的判别条件。比如：括号位置的候选词语，需要是动词或介词，而且要同时能跟“鲁迅”和“作品”发生述宾结构关系。10 个候选词语中，“收藏、拍卖”虽是动词，但不能跟“鲁迅”发生述宾结构关系；“青年”“创举”等是名词，都不符合替换要求。只有“批、评、关于、赞美”这几个词是动词或介词，能跟“鲁迅”和“作品”组成述宾结构，符合替换要求。数据驱动派的做法是计算每个候选词在括号位置的分布概率，计算依据来自训练海量语料所获得的“经验”，即评估每个候选词进入到括号位置后所形成词组（包括不同结构关系）合法的概率，以及词组中各词语之间结构关系跟“讽刺鲁迅的作品”中各词语间关系的相似程度。

知识驱动的计算机语言处理程序，可以在微观层面描述计算机运行中所依赖的知识，以及在运行流程中知识所起作用的细节情况。而数据驱动的神经网络模型，很难在微观层面说明亿万参数激活的传导过程跟最终输出答案之间的因果联系。句中每个字符在运算过程中都转成了向量形式，而每个向量在大语言模型中都可能高达上万维，这些维度共同表达了字符的意义、用法、位置、语境等综合信息，在如此复杂的向量空间中，具体每个维度如何影响模型对句子的生成和语义理解，难以跟踪得到因果解释。采用数据驱动的方法，计算机不是直接学习专业语言学知识，而是学习人类的言语实践经验。人工智能发展到今天，数据驱动派已经实现了以大语言模型（Large Language Model）为代表的多任务自然语言处理系统，在许多自然语言任务场景中表现出色。但人工神经网络的可解释性差这一问题仍很突出。反观知识驱动派的系统，则没有这方面的缺点。其内部知识是人可懂的表征形式，语言处理过程是可解释的。不过，直接依靠人类知识构建计算机自然语言处理系统，在知识的规模和质量方面面临巨大的挑战，往往难以实现跨任务的灵活的语言处理能力。乔姆斯基曾把语言学研究的目标概括为三个层级：观察充分性（observational adequacy）、描写充分性（descriptive adequacy）、解释充分性（explanatory adequacy）。以上文所举异形同义现象来讲，观察充分性就是要全面地收集主宾可换位的例句，特别是像“回归”这样的边缘例子。后两个要求是要构建理论模型，对语言材料进行系统高效地分

类整理和描写,并进一步在产生动因、表达功能等方面做出解释,比如对主宾换位句,就可以通过动词论元结构、名词物性结构、构式整体语义等多个理论模型来阐释其机理。对数据驱动派来讲,神经网络训练材料来自整个互联网海量数据(既包括自然语言,也包括程序代码,各领域行业文献等人工语言数据,甚至还包括图片、音频、视频等多模态数据),在观察充分性上远超人类个体语言研究者所能掌握的数据范围。但在描写充分性和解释充分性上,显然存在欠缺。对知识驱动派来讲,关于人类语言的研究,这三个充分性的目标达成了多少,往往是用对局部语言现象的研究去衡量,如果从言语数据的总量,在语言宏观层面来审视,无疑也是远远不够的。知识驱动和数据驱动,各自都是既有所长,亦有所短。两条路线的探索都还会继续,未来也必然会出现融合两派优点的更加先进的理论模型。

2019年英国银行发行了纪念图灵的50英镑纸币,上面印着图灵在1949年评论当时英国曼彻斯特大学“机械大脑”(mechanical mind)研究成果的一句话:**This is only a foretaste of what is to come, and only the shadow of what is going to be.**(毫芒见鸿影,万卷藏序章)。70多年过去,无论是数据驱动还是知识驱动的计算机模型,模仿人类语言的能力都取得了长足的进步,甚至有人说计算机已经通过图灵测试。但用“测试”来描述人机关系,其实不如图灵原创的“游戏”。计算机对人类语言的“模仿游戏”理应继续下去,直到我们真的能清晰地回答“什么是语言”这个关乎人何以为人的本质问题。

参考资料

- [1] 北京大学中文系应用语言学专业官网: <https://ccl.pku.edu.cn/all/>
- [2] 詹卫东、王佳骏, 2022, 从感知智能跃升到认知智能——计算语言学的机遇与挑战, 北京大学中国语言文学系编《四海文心——我与北大中文系》北京大学出版社, 2022年1月出版。下册, 634-659页。
- [3] 方立、李谷城等译, 1993, 《语言要略》外语教学与研究出版社。
- [4] Dwight Bolinger, [1968] 1975, *Aspects of Language*, Thomson Learning; 2nd edition. New York: Harcourt, Brace, Jovanovich.
- [5] Alan Turing, 1950, Computing Machinery And Intelligence. *Mind*, 59, 433-460.
- [6] Terry Winograd, 1971, Procedures As a Representation For Data In A Computer Program For Understanding Natural Language, PhD Thesis, Massachusetts Institute of Technology, January 1971.
- [7] Terry Winograd, 1983, *Language As A Cognitive Process*, Volumn 1 Syntax, Addison-Wesley Publishing Company.
- [8] Noam Chomsky, 1957, *Syntactic Structure*. The Hague: Mouton.
- [9] Noam Chomsky, [1968] 2006, *Language and Mind*, Cambridge University Press.

补记:北京大学中文系应用语言学本科专业成立于2001年,是在中文系传统语言学专业基础上为适应新时代学科发展需要做出的重大创新。有关该专业情况的详细介绍见本文所附参考资料[1]、[2]。本文所谈语言学和人工智能的过往、现状及研究目标,是结合近几年生成式人工智能,特别是大语言模型技术的飞速进步,对应用语言学内涵的新的思考,是对有关应用语言学专业介绍的一个补充和深化。

致谢:北京外国语大学陈国华教授帮助修正了本文初稿中的译名不当并提出宝贵修改意见,特此致谢。文中谬误,概由作者本人负责。