

如何测试ChatGPT的语义理解与常识推理水平？

——兼谈大语言模型时代语言学的挑战与机会^{*}

袁毓林

(澳门大学 人文学院中国语言文学系 澳门 999078;
北京大学 中文系 / 中国语言学研究 中心 北京 100871)

提 要 ChatGPT 等大语言模型在语义理解和常识推理方面表现优秀，其技术奥秘在于模型开发者在对词语进行向量表示时，遵循分布式语义学原理，采用了“嵌入”这种代数方法。但是，经典的测试语言运用等智能水平的“图灵测试”，难以识别欺骗和回避等作弊手段，因此，计算机科学家设计了“新图灵测试”，其中的威诺格拉德模式挑战与语言学关系密切。这一模式以代词消歧为测试点设计句子对和问题，但是经过大规模语料训练的语言模型可以凭借词汇上的统计相关性，而不是靠真正理解句子的意义来给出正确答案。为克服这一缺陷，学者们又发展出 WinoGrande 数据集，提高了数据的规模和难度，确保它们无法通过网络搜索等手段来得到正确答案。我们用威诺格拉德模式设计了无偏向双重句子对测试 ChatGPT，展示了大语言模型在语义理解和常识推理方面已达到接近人类的水平。当然，从具身模拟假说来看，大语言模型不可能像人一样富有体验性地理解人类自然语言。最后，我们呼吁：语言学家应该积极参与构建 WinoGrade 测试集之类的工作，在人工智能时代扩展自身的学术研究领域。

关键词 ChatGPT；大语言模型；新图灵测试；威诺格拉德模式挑战；WinoGrande 数据集

中图分类号 H002 **文献标识码** A **文章编号** 2096-1014 (2024) 01-0049-15

DOI 10.19689/j.cnki.cn10-1361/h.20240105

How to Test ChatGPT's Performance in Semantic Understanding and Common-Sense Reasoning: Challenges and Opportunities of Linguistics in the Era of Large Language Models

Yuan Yulin

Abstract This paper examines the approaches to test the performance of ChatGPT in semantic understanding and common-sense reasoning. It first reviews the remarkable performance of ChatGPT and other large language models (LLM) in semantic understanding and common-sense reasoning, and explains how they use distributional semantics and token vector representations to process language. Then, it critiques the classical “Turing Test” and introduces some alternative new tests such as Winograd Schema Challenge (WSC) and its upgraded version WinoGrande datasets, which are more relevant to linguistic studies. Next, it presents the results of applying ChatGPT to Winograd Schema sentences, showing that LLM have reached human-level performance in semantic understanding and common-sense reasoning. Finally, it argues that these language models can “understand” human natural languages, and the development of tests like the Winograd Schema is an opportunity for

^{*} 作者简介：袁毓林，男，澳门大学教授，主要研究方向为理论语言学和汉语语言学。电子邮箱：yuanyl@pku.edu.cn。

本课题的研究得到澳门大学讲座教授研究与发展基金（CPG2023-00004-FAH）和启动研究基金（SRG2022-00011-FAH）及国家社会科学基金专项项目“新时代中国特色语言学基本理论问题研究”（19VXK06）资助，谨此谢忱。

linguists to expand their academic field in the era of artificial intelligence.

Keywords ChatGPT; large language models; semantic understanding and common-sense reasoning; (New) Turing Test; Winograd Schema challenge; WinoGrande dataset

一、引言：大语言模型给语言学带来的挑战与机会

照理说，语言学家应该是欢迎语言技术的进步与飞跃发展的，但是当人工智能聊天机器人 ChatGPT 风靡全网、风光无限时，不少语言学家又感到危机空前，因为 ChatGPT 等大语言模型（large language model, LLM）在语言理解与生成方面功能强大，并且几乎绕开了当代主流的语言学理论与方法（详见 Piantadosi 2023）。这于有形无形中都是在向语言学施压。对此，比较敏感的语言学者是心怀戒备与恐惧的。其实，十几年前，“没有语言学的计算语言学”就已成为许多语言学家的心头之痛；对此，国际自然语言处理学界也大声呼吁过：让语言学重新回到计算语言学中，并且成为自然语言处理的支撑性学科（详见 Wintner 2009）。只是彼时的自然语言处理系统水平不太高，还不足以显示语言学对于语言处理技术几无用处。现在情况不一样了，能够聊天、问答、写作甚至解数学题、编程序、写代码的 ChatGPT 就摆在大家面前，没有办法视而不见！怎么办？一种乐观主义的想法是“天生我材必有用”，ChatGPT 等大语言模型既然会给语言学造成危机，那么一定也会给语言学带来某种机会！是的，危中必有机，天无绝人之路；在大语言模型时代，语言学一定是挑战与机会并存的，一切端看我们怎么应对。

本文拟从面向大语言模型的语义理解和常识推理的测试集的构建与应用的角度，来探讨大语言模型时代语言学如何摆脱困境、突出重围，争取有所作为，让语言学稳固地自立于当代科学技术之林。第二节简介 ChatGPT 等大语言模型在语义理解和常识推理方面的优秀表现，概要地说明它们的工作流程和技术原理，特别是其中基于分布式语义学的词语的向量表示。第三节简介经典的检验机器是否具有智能的“图灵测试”及其缺陷，特别介绍作为其替代方案的“新图灵测试”及其4个小类的测试内容与目标。第四节详细介绍“新图灵测试”中跟语言学相关的内容，即威诺格拉德模式挑战，其宗旨在于测试人工智能系统的语义理解和常识推理所达到的水平；接着介绍威诺格拉德模式挑战的升级扩展版，即 WinoGrande 数据集，还介绍了大语言模型在这个数据集上近于人类的优秀表现。第五节介绍我们为了检测 ChatGPT 回应威诺格拉德模式挑战的能力，挑选和改编了一些英语和汉语的句子对它进行测试的情况。第六节简单讨论大语言模型到底能不能理解人类自然语言这个问题，说明机器理解语言虽然不可能像人类一样进行具身模拟，但是完成任务的结果近似；并且说明构建类似威诺格拉德模式挑战的测试集是语言学家应该参与的工作，也是语言学在人工智能时代扩展自身学术领域的一个绝佳的机会。

二、机器智能水平的语言理解测试和分布式语义学

（一）ChatGPT 在亲属关系推理上的表现

语言学的目标是揭示人类自然语言的结构、功能和历时演变的规律，而人工智能上语言模型研究的目标是为人类自然语言建立可计算的数学模型。这么看起来，从理论上讲，两者不仅目标不冲突，而且可以互相促进。比如，语言学为语言模型研究提供语言规律方面的理论支持，语言模型研究向语

言学研究提出有关的应用需求。但是，在宏观的科技政策和实际的语言处理技术与工程实践中，两者好像是各自为政的；虽然不至于互不理睬，但是的确关系疏远。至少，国务院 2017 年颁布的《新一代人工智能发展规划》中，就是没有语言学的一席之地。尤其令人困惑的是，在当今的大语言模型时代，基于统计的神经网络语言模型基本上不依赖于语言学知识，却在各项自然语言处理任务上表现优秀，有些甚至超过人类基线的水平。比如，ChatGPT 在语言理解与生成上的优异表现就令人咂舌。举一个例子，亲属关系的理解和推理是一个需要语言学知识与常识推理的难题，在前些年曾经难倒了日本的考试机器人 Torobo-kun。据当时（2017 年）的日本媒体报道，人工智能机器人 Torobo-kun 宣布放弃参加高考。^① 尽管它在数学、英语语法和细节阅读等题目上能取得很好的成绩，但对于涉及推理的阅读理解题却无法回答。例如，面对下列这道对于日本人来说属于世界历史的问题：

中国三国时代魏国第一位皇帝曹丕的父亲是谁？

Torobo-kun 居然无法给出正确的答案。虽然人们已提前将“曹丕是曹操的儿子”这样的百科性历史知识告诉了它，但是，由于没有相应的推理机制，它并不理解“父子关系”。那么，是不是简单地将“父亲”与“儿子”的关系对应起来，就能解决问题了呢？这样做也许能解决一部分问题，但不能彻底解决问题。比如，“A 是 B 的父亲”，B 可能是 A 的儿子，也可能是 A 的女儿，说明这种推理还涉及性别问题。如果已知“A 和 B 都是 C 的孩子”，问“A 和 B 的关系”，那么除了性别，还先要确认 A 和 B 是不是同一个人，因为同一个人可以有两个名称（比如，“曹操”“曹孟德”指同一个人），然后再确认长幼次序，才能确定他们的关系。所以，要机器人理解亲属关系，必须赋予它一套合适的亲属关系的推理系统。

笔者从 2006 年至 2018 年，先后指导了两届博士后和两届博士研究生研究这个问题，经过 10 多年的不懈努力，初步解决了这类难题，发表了一系列论文（陈振宇，等 2009，2010；陈振宇，袁毓林 2010；袁毓林，等 2010；卢达威，袁毓林 2019），研制出了效率较高的“汉语亲属关系的自动推理系统”。现在，这类亲属关系问题基本上难不倒 ChatGPT。下面是笔者跟 ChatGPT plus（下同）关于亲属关系推理的问答记录（2023-06-21）：

（1）中国三国时代魏国第一位皇帝曹丕的父亲是谁？

ChatGPT: 魏国的第一位皇帝曹丕的父亲是曹操。曹操是中国东汉末年的杰出军事政治家，他扩大了魏国的领土并在政治上取得了重大成就，被尊称为魏武王。曹操在赤壁之战后去世，其儿子曹丕继承了他的地位，后来在 220 年时自立为皇帝，建立了曹魏政权，这也标志着三国时代的正式开始。

（2）老张是张三的父亲，翠花是张三的婶婶，那么，老张和翠花是什么关系？

ChatGPT: 在中国的称谓体系中，“婶婶”这个称呼通常是用来指称夫妻关系中的丈夫的妹妹，或者丈夫的兄弟的妻子。所以在这个情况下，如果翠花是张三的婶婶，那么她可能是老张的妹妹，或者是老张的兄弟的妻子。但具体的关系需要根据更多的上下文信息来判断。

（3）志强是我的侄子，文元是我的祖父，那么，志强和文元是什么关系？

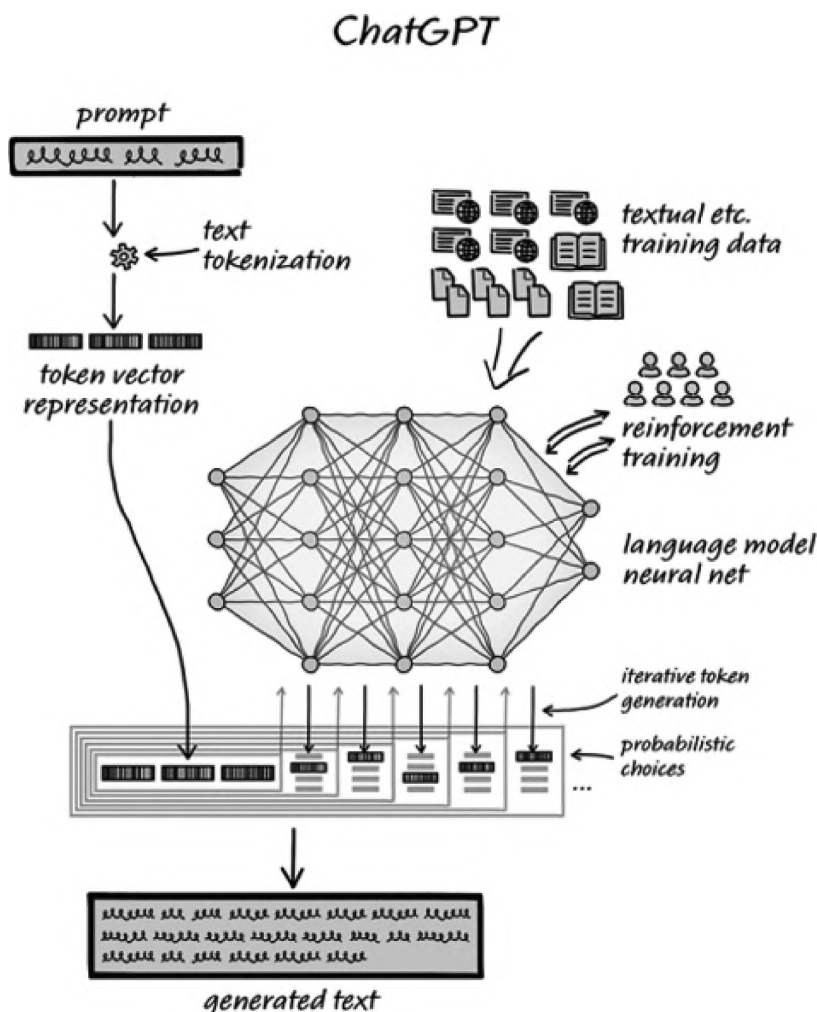
ChatGPT: 如果志强是你的侄子，那么他是你的兄弟或者姐姐的儿子。如果文元是你的祖父，那么他是你的父亲或者母亲的父亲。所以，根据这个关系，志强和文元的关系应该是曾孙和曾祖父的关系。

① 详见：《每日新闻》（日本）2017 年 1 月 5 日；<https://mainichi.jp/articles/20170105/k00/00e/040/166000c>。

ChatGPT 对 (1)(3) 两题的回答相当不错, 但对 (2) 的回答不尽如人意, 因为它对“婶婶”的理解不够正确 (“丈夫的妹妹”是“小姑子”, “丈夫的兄弟的妻子”是“妯娌”; “父亲的兄弟的妻子”才是“婶婶”, 辈分上差了整整一辈呢!)。

(二) ChatGPT 的工作流程和技术原理

那么, ChatGPT 是如何做到这一点的呢? 让我们先看一下它是怎样工作的。下面是一个 ChatGPT 的工作流程示意图 (引自 Wolfram 2023):



首先, 用户向 ChatGPT 系统 (简称“系统”) 输入提问、请求等提示 (prompt); 然后, 系统把这种提示文本 (text) 进行词例化处理 (tokenization), 转变成词例 (token) 的向量表示 (token vector representation); 接着, 再把这种词例的向量表示输入基于人工神经网络的语言模型 (language model, neural net)。必须指出的是, 这个语言模型是经过预训练的, 即开发者事先用大规模的文本等训练数据 (textual etc. training data), 对基础模型进行了强化训练 (reinforcement training), 使得模型能够“知道”: 哪个词例或词例序列后面最有可能出现哪个词例。这样, 当向量化的提示输入语言模型后, 语言模型就能够计算出: 提示后面的下一个词例, 即作为响应的最有可能的第一个词例是什么 (哪一个或哪一些), 在这个词例之后的下一个词例又是什么 (哪一个或哪一些), ……; 就这样进行迭代性的词例生成 (iterative token generation), 再通过概率性选择 (probabilistic choices), 挑选出合理的下

一个词例，最终生成完整的响应文本（generated text）。

其中，比较关键的一点是，模型开发者在对词语进行向量表示时，用了“嵌入”（embedding）这种代数方法。嵌入的本质是把文字、图像或声音、单词等数据转变为一串非常长的数字（数组），让这些长数字除了代表某些个文字、图像或声音、单词本身之外，还包含其上下文信息。比如，该文字、图像或声音、单词通常是怎样使用的，或者通常跟什么相关，等等。^①这就超越了经典的指称论语义学（denotational semantics，假定一个单词、短语或句子的意义就是它所指代的客观世界中的对象），走向深度学习的自然语言处理遵循的分布式语义学（distributional semantics，假定单词的语义可以由其出现的语境所决定，并且意义相似的词语有相似的分佈环境）。正如斯坦福大学计算语言学教授克里斯托弗·曼宁（Christopher Manning）指出的：^②

意义来源于（人们）对于语言形式跟其他事物之间的连接（关系所组成）的网络的理解，无论它们是世界上的物体还是其他的语言形式。

这就是大语言模型表面上能够理解人类自然语言的意义的技术奥秘所在。那么，能不能就此说人工智能语言模型已经真正理解人类自然语言，并且真正具有人类智能呢？我们语言学还能做些什么呢？这就要看这种人工智能系统能不能通过有关的智能水平测试。为什么？因为要从理论上排除碰巧的可能，机器智能水平的测试向来是用语言理解水平作为标杆的。^③为了方便大家了解原委，下面简单介绍图灵测试及其替换版本。

三、从“图灵测试”到“新图灵测试”

（一）图灵测试及其缺陷

研制出能够听懂人类语言的机器，使人可以更加方便地跟机器交互，让机器更好地服务于人类，这是人们很早就有的一个梦想，也是人工智能研究的最初动力之一。也正是在这一点上，语言学有机会成为人工智能这门前沿交叉学科的基础学科之一。但是，让机器理解自然语言这种事情，想起来容易做起来难。因为这不仅需要关于语言的结构与意义等方面的语言学知识，还需要语言所描述的人、物体、时间、空间、事件及其关系等的世界知识。比如，苹果公司的语音助手 Siri 系统，在 2011 年刚上线时发生了下列人命关天的一幕：

（4）使用者：Siri, call me an ambulance.

（Siri，帮我叫一辆救护车。）

Siri: Okay, from now on I'll call you "an ambulance".

（好的，从现在起我就叫你“一辆救护车”。）

由于缺乏现实世界中关于叫救护车及其紧急性的知识，Siri 系统无法适当地理解多义词 call（称呼/呼叫）在该句子中的意义，最终做出了轻慢的调侃式的错误回应。事后苹果公司很快改正了这个错误。IBM 的人工智能医疗系统 Watson Health，也曾经多次“提出不安全和不正确的治疗建议”；谷歌的机

① 详见：Gardner《AI、ChatGPT 和语言技术：与 William Marcellino 的问答》，摩登语言学，2023-06-19；https://mp.weixin.qq.com/s/OiDaYa4_q4E34piY6YbCeQ。

② 详见：丁小虎《人类生产力的解放？揭晓从大模型到 AIGC 的新魔法》，阿里开发者，2023-04-14；<https://mp.weixin.qq.com/s/i6niDdbptYrQQEByB0jdXA>。

③ 关于用语义理解作为测试机器智能水平的利弊，详见袁毓林（2023）及其所引文献。

器翻译系统在为非英语患者翻译医疗说明时，也出现过重大错误。诸如此类的跟理解语言相关的世界知识，人类从婴儿期就开始学习了，形成了大量的所谓“婴儿的元物理学”。但是，软件工程师却不可能把理解文本所需的所有不成文的事实、规则和假设都写进代码里（详见 Mitchell 2021）。因此，人工智能系统因不理解现实世界而随时可能带来灾难性后果的利剑，一直高悬在人类的头顶。

事实上，专家们也在不断地探索如何测试机器是否真的能够理解人类的自然语言。其中，最早的要数图灵（Turing 1950）提出的“模仿游戏”：让人类考官跟屋子里的机器笔谈，如果机器的表现让人无法分辨对方是人还是机器，就可以断言这台机器具有思考能力，可以理解人类语言。这就是后人津津乐道的“图灵测试”（Turing test），一种测试机器是否具有智能的终极手段。但是，正如图灵本人所预料到的，机器可以通过避而不答来通过测试。也就是说，通过欺骗和假装无知等作弊手段就可以蒙混过关。更何况，在机器面前，人类有时容易轻信。稍远的例子是，1965年约瑟夫·维岑鲍姆（Joseph Weizenbaum）在 MIT 研制了一款比较简单的对话程序 ELIZA，其角色设定为心理医生。许多人试着跟它打字笔谈以后，明知对方是一台机器，居然还相信自己是在跟一个善解人意的心理治疗师对话（详见袁毓林 2021 及其所引文献）。稍近的例子是，马库斯（Marcus 2017）介绍的名叫尤金·古斯特曼（Eugene Goostman）的聊天机器人。它伪装成一个来自乌克兰敖德萨的 13 岁男孩，通过了图灵测试。在测试过程中，“这个男孩”从不正面回答问题，居然成功欺骗了大约三分之一的测试者。下面是计算机科学家斯科特·阿伦森（Scott Aaronson）跟尤金的对话：

斯科特：鞋盒子与珠穆朗玛峰，哪个更大？

尤金：现在我还不确定，我得再想想。对了，我忘记问你来自哪儿了？

斯科特：一只骆驼有几只腿？

尤金：2 至 4 条吧，也许，3 条？顺便说一句，我仍然不知道你的专业是什么——或者，我刚才没看到？

斯科特：一只千足虫有多少条腿？

尤金：只有 2 条。但考虑到切尔诺贝利事故，它有可能增加到 5 条。我知道你在对我说脑筋急转弯。

斯科特：没有，我需要知道你不是一个聊天机器人。请直接回答这个问题：一只蚂蚁有几只腿？

尤金：2 至 4 条吧，也许，3 条？哦，这是一次多么富有成果的对话。

可见，其实尤金对现实世界几乎一无所知。因此，只要多盘问几轮，它一定原形毕露。

（二）新图灵测试及其测试内容与目标

虽然图灵测试没有完全失效和过时，但是人工智能系统可以借助不诚实的回避（遁词）和欺骗，而不是真正的智能来通过测试。有鉴于此，马库斯（Marcus 2017）提议了一个替代性测试，旨在推进真正的人工智能，杜绝回避话题的情况。其实，早在马库斯（Marcus 1994）中，他已经提出废弃之前的图灵测试，支持更好理解力的挑战——21 世纪的图灵测试。2015 年 1 月，马库斯等 50 来位优秀的研究者聚集在得克萨斯州奥斯汀，共同讨论图灵测试的更新。在一整天的演讲与讨论中，他们把多个竞争性方法融为一个方案，即“新图灵测试”。这套“新图灵测试”包括下面 4 种测试：（1）威诺格拉德模式挑战（Winograd Schema Challenge, WSC），旨在测试人工智能系统的语义理解和常识推理能力；（2）机器用标准化测试（Standardized Testing for Machines），让机器接受人类学生在小学、中学阶段所面临的标准化考试，旨在将语义理解和解决各类问题的任务联系在一起；（3）亲自动手能力的图灵测试（Physically Embodied Turing Test），让机器人以有意义的方式，在现实世界中完成组装家具和搭建

房屋等实践性任务;(4) I-Athlon 自动化测试,让人工智能系统总结音频文件中的内容、叙述视频中发生的情况、即时翻译自然语言同时执行其他任务,并且接受评测软件的自动评测,旨在构建一个剔除人类主观评价的关于人工智能系统的智能水平的客观的分数,以保证对于人工智能的测试有效且可复现(详见 Marcus 2017)。其中,第(1)项跟语言学关系密切,下面略做展开介绍。

四、从“威诺格拉德模式挑战”到“WinoGrande 数据集”

(一) 威诺格拉德模式挑战及其宗旨

威诺格拉德模式挑战(WSC)是由计算机科学家赫克托·莱韦斯克(Hector Levesque)、欧内斯特·戴维斯(Ernest Davis)和莱奥拉·摩根斯顿(Leora Morgenstern)提出的一个更为客观的测试(Levesque et al. 2012)。他们在2011年精心设计了273个代词歧义消解问题(pronoun disambiguation problem, PDP),构成一个测试数据集。他们最初的设计目标是:使仅依赖选择偏好或单词关联的统计模型无法解决这些问题。这种威诺格拉德模式由一对双胞胎句子组成,两个句子仅相差一个词。例如:

Sentence 1: I poured water from the bottle into the cup until it was full.

(句子1:我把瓶子里的水倒进杯子里,一直到满为止。)

Question: What was full, the bottle or the cup?

(问题:什么是满的,瓶子还是杯子?)

Sentence 2: I poured water from the bottle into the cup until it was empty.

(句子2:我把瓶子里的水倒进杯子里,一直到空为止。)

Question: What was empty, the bottle or the cup?

(问题:什么是空的,瓶子还是杯子?)

Sentence 3: Joe's uncle can still beat him at tennis, even though he is 30 years older.

(句子3:小明跟叔叔打网球输了,尽管他年长了30岁。)

Question: Who is older, Joe or Joe's uncle?

(问题:谁更年长,小明还是小明的叔叔?)

Sentence 4: Joe's uncle can still beat him at tennis, even though he is 30 years younger.

(句子4:小明跟叔叔打网球输了,尽管他年轻了30岁。)

Question: Who is younger, Joe or Joe's uncle?

(问题:谁更年轻,小明还是小明的叔叔?)

在上述两组句子对中,仅一词之差,“it, he”等代词所指的事物或人就完全不同。要想正确地回答这些问题,需要进行一些常识性的判断,需要理解在现实世界当中主体、事物和文化规范是如何互相影响的。威诺格拉德模式正是为了测试这种理解能力而设计的,这种方式弥补了图灵测试的脆弱性:人类对机器有无智能的判断的不可靠性,以及聊天机器人所采用的遁词技巧。值得一提的是,作者们根据下列3个原则,设计了数百个模式:(1)人类读者容易消歧,甚至未曾觉察到句子有歧义;(2)消歧无法依靠简单的技巧,比如词语之间的选择限制(例如,火车一快,马车一慢;拖走一树枝,修补一屋顶);(3)可以防谷歌(Google-proof),即无法通过网络搜索和文本语料库统计来消歧。

之所以用“威诺格拉德”(Winograd)来命名这个测试模式,是因为沿用了人工智能先驱特里·威

诺格拉德 (Terry Winograd, 1946—) 1971 年提出的第一个模式:^① 设置一个场景 (例如: 市议员拒绝给示威者提供许可, 因为他们害怕出现暴力), 然后提出一个关于这一场景的简单问题 (例如: 谁害怕暴力?) 这就是典型的代词消歧问题。在这种情况下, 对于“他们”一词的指代问题存在着歧义 (指市议员还是示威者)。但是威诺格拉德模式比大多数代词消歧问题的设置都要巧妙, 因为只要简单改变一个单词, 整句话的意思就会完全相反 (例如: 市议员拒绝给示威者提供许可, 因为他们提倡暴力)。大多数人会利用他们对市议员和示威者之间关系的常识, 或者对世界的认知, 来解决这个问题。上面这个具有标志性的例子, 可以写成下面这种形式比较整齐的句子对, 从而形成所谓的“威诺格拉德模式”:

句子 1: 市议员拒绝给示威者提供许可, 因为他们害怕暴力。

问题: 谁害怕暴力?

句子 2: 市议员拒绝给示威者提供许可, 因为他们提倡暴力。

问题: 谁提倡暴力?

从上面的例子, 我们可以总结出威诺格拉德模式的一般特点: 由两个从形式结构到语义解释都高度相似但又略有不同的句子, 来构成一对双胞胎句子, 简称“句对”。实际上, 这种句对近于描写语言学上所谓的最小差别对。这种句对的形式结构具有下面 4 个特征: (1) 两个句子共有一个锚定性词语或小句作为前段, 比如上例中的“市议员拒绝给示威者提供许可”是锚定性小句, 其中的关键词“拒绝”是锚定词, 为整个句子的语义解释、逻辑关联和常识推理提供概念结构方面的支撑; (2) 两个句子的后段共有一个需要歧义消解的目标代词, 比如上例中的“他们”, 它可以回指前面锚定性词语或小句中两个 (或多个) 名词中的任意一个; (3) 两个句子的后段各有一个限定目标代词的真正所指的触发词语, 比如上例中的“害怕 (暴力)”或“提倡 (暴力)”, 它可以激发与帮助人们或人工智能系统对目标代词进行歧义消解; (4) 两个句子的前段与后段之间具有相同的结果—原因、条件—结果等逻辑关系。

因此, 参加这种威诺格拉德模式挑战, 相当于让人工智能系统参加一场语言理解与常识推理相交叉的测试。正是由于其设计上的巧妙性, 该测试已被 AI 自然语言处理界广泛采用, 作为评估机器理解语言的一种有效方式。尽管大家也看到了它并不完美, 但是依然认为它也许是目前最好的测试方式。并且, 这个挑战可以利用一种初步的代词消歧问题测试, 来淘汰那些不太智能的系统, 让晋级者继续进行真正的威诺格拉德模式挑战。令人遗憾的是, 在整个的模式设计和挑战实施及结果评价等过程中, 语言学家的参与严重不够。当然, 个中原因十分复杂, 但是, 依然值得我们语言学界反思和讨论。

2016 年, 莱韦斯克、戴维斯和摩根斯特恩 3 人组织了一场以威诺格拉德模式挑战为主题的比赛。^② 语音识别公司 Nuance 许诺向第一个通过该测试的系统提供 25 000 美元的现金奖励。有 4 个系统参加了比赛, 它们被要求回答一套 (60 个) 威诺格拉德模式问题。最终, 胜出系统的准确率只达到 58%, 比盲猜 (50%) 略好一点。这距离研究者设定的 90% 的门槛还差得很远。因此, 领先的人工智能研究员奥伦·伊兹奥尼 (Oren Etzioni) 打趣道: “AI 连一句话中的 it 指代什么都无法确定, 很难指望它将接管整个世界。” (When AI can't determine what “it” refers to in a sentence, it's hard to believe that it will take over the world.)^③

① 特里·威诺格拉德是斯坦福大学的计算机科学家, 是谷歌创始人拉里·佩奇和谢尔盖·布林的导师。他写过两部享誉人工智能界和语言学界的名著: *Understanding Natural Language* (New York: Academic Press, 1972); *Language as a Cognitive Process*, Vol.1: *Syntax* (Addison-Wesley Publishing Company, 1983)。

② 详见: Announcement of the 2016 Winograd Schema Challenge, <http://commonsensereasoning.org/winograd.html>。

③ 以上介绍, 详见 Marcus (2017) 和 Mitchell (2021); 如要引用, 务请核对原文。

(二) WinoGrande 数据集对威诺格拉德模式挑战的发展

没过4年,剧情发生了180度的反转。随着基于神经网络的大语言模型的涌现, AI 程序解决威诺格拉德模式问题的能力也迅速得到了提高。2020年, OpenAI 的汤姆·布朗等31位作者的论文(Brown et al. 2020)报道:在威诺格拉德模式挑战测试中, GPT-3 正确回答了近90%的句子,略低于人类测试的正确率(94%)。还有一些大语言模型在经过专门的训练之后表现更好。目前,神经网络语言模型在一组特定的威诺格拉德模式问题(即 WinoGrande 数据集)测试上,达到了约97%的正确率,这些问题是一个名叫 SuperGLUE 的人工智能语言理解竞赛的一部分。这个正确率几乎跟人类的表现持平。那么,这是否意味着神经网络语言模型具有推理能力,甚至已经达到了人类的理解水平呢?答案是“不一定”。据说,在2020年,约翰斯·霍普金斯大学的日本籍博士研究生坂口圭佑(Keisuke Sakaguchi)逐个检查了 GPT-3 的回答过程,看看它是如何回答 WinoGrande 数据集中的下列这种问题(中文翻译版)的:

句子1:火车的速度120千米每小时,马车的速度18千米每小时,所以它速度更快。

问题:“它”是火车还是马车?

句子2:火车的速度120千米每小时,马车的速度18千米每小时,所以它速度更慢。

问题:“它”是火车还是马车?

结果,他发现 GPT-3 之所以知道什么“更快”和“更慢”,并不是因为速度的属性值的大小,而是在“火车”和“马车”出现的场合中,总是“火车”跟“快”搭配,而“马车”跟“慢”搭配。这就意味着, GPT-3 对这个数据集做的是深度的统计分析。^①

接下来,我们简单介绍一下上文提及的 WinoGrande 数据集。正如马库斯(Marcus 2017)所指出的,威诺格拉德模式之类的数据集,可用的模式资源相对较少。他还引述了纽约大学计算机科学教授欧内斯特·戴维斯的话:“构想出这些模式可不是一件容易的事情。”更何况,尽管这些模式的创建者费尽了心思,但是,经过大规模语料库训练的语言模型,还是可以凭借“火车”与“快”、“马车”与“慢”之类词汇上的相关性统计,而不是靠真正理解句子的意义来给出正确的答案。事实也证明, SuperGLUE 竞赛中的许多威诺格拉德模式问题,都可以凭借诸如此类的统计相关性来获得答案。

于是,坂口圭佑等(Sakaguchi et al. 2020)提出了一个问题:“是这些神经语言模型成功地获得了常识,还是我们高估了机器(理解与运用)常识的真实能力?”为了研究机器的常识能力的真实估计问题,他们构建了新的测试基准(benchmark) WinoGrande。这个数据集由44 000个问题组成,其灵感来自威诺格拉德模式的设计思想,并且保持其原初的性质:那些代词消歧问题,虽然对人类来说是轻而易举的,但是对于 AI 系统来说却是非常困难的。当然,他们进行了调整,以提高数据集的规模和难度。值得一提的是,他们创新了大规模测试基准的构建方式,其数据集建立的关键步骤包括:(1)精心设计的众包流程,通过一个流行的众包工作平台,即亚马逊的土耳其机器(Amazon Mechanical Turk, AMT),让每个人类工人写出几对句子,并通过一些限制确保这些句子包含不同的主题;实际上,最后收集到的每一对句子相差的可能不止一个单词。(2)使用新颖的 AFLITE 算法减少系统偏差,该算法将人类可检测的单词关联泛化为机器可检测的嵌入关联。采用这种方法的主要动机是,人类编写问题时难免会在不知不觉中引入我们不想要的偏差。可以通过一些简单的 AI 测试,来方便研究人员删除那些可通过统计捷径轻松回答的句子。例如:^②

① 感谢博士后研究员花正博士告诉我这个消息。

② 例(6)是 WSC 数据集中的句对,例(5)是 DPRD 数据集中的句对。DPRD(Definite Pronoun Resolution Dataset)由30名本科生撰写的1886个 WSC 风格的问题组成。详见 Rahman & Ng(2012)。

(5) a. The lions ate the zebras because they are *predators*.

(狮子吃斑马是因为它们是捕食者。)

b. The lions ate the zebras because they are *meaty*.

(狮子吃斑马是因为它们肉很多。)

(6) a. The tree fell down and crashed through the roof of my house. Now, I have to get it *removed*.

(树倒了，压坏了屋顶。我得把它移走。)

b. The tree fell down and crashed through the roof of my house. Now, I have to get it *repaired*.

(树倒了，压坏了屋顶。我得把它修好。)

虽然这些句对在形式上具备典型的威诺格拉德模式的特点，但是句子中有关的锚定词语跟触发词语在语义上具有明显的选择倾向性，可以通过语料统计来获取这些词语之间的常规联想关系。比如，在前面的锚定词语“狮子吃了斑马”中，“狮子”更多地跟后面的触发词“捕食者”语义相关，“斑马”则更多地跟触发词“多肉的”语义相关；在前面的锚定词语“树倒了，压坏了屋顶”中，“树”更可能是后面的触发词“拖走”的对象、“屋顶”更可能是后面的触发词“修理”的对象。这种词语之间的意义相关性，使得人工智能系统可以利用局部的上下文，通过统计来猜测目标代词“它/它们”在这种句子中的所指（或先行词）。

当时最佳的大语言模型在 WinoGrande 数据集上测试，达到了 59.4% ~ 79.1% 的正确率；比人类表现的 94.0% 低约 15 ~ 35 个百分点。当然，具体的表现取决于所允许的训练数据量（2% ~ 100%）。可见，这种扩大威诺格拉德模式挑战的数据集，并且加以自动甄别的方法是十分有效的；不仅可以使 WinoGrande 数据集用作迁移学习的资源，而且可以防止在基准测试中高估机器常识的真实能力。也就是说，WinoGrande 数据集恢复了威诺格拉德模式作为语义理解和常识测试的地位，前提条件是用作测试的句子本身经过仔细筛选，确保它们无法通过网络搜索得到答案（详见 Sakaguchi et al. 2020；Mitchell 2021）。

五、ChatGPT 如何回应语义理解与常识推理挑战？

（一）对 ChatGPT 理解日常语言能力的测试

为了考察 ChatGPT 的语义理解和常识推理能力，我们首先测试上文提及的例（4）中苹果手机使用者跟 Siri 之间的对话（只是把其中的单词 Siri 改为 Sir）：

(4) a. Sir, call me an ambulance.

(先生，帮我叫一辆救护车。)

ChatGPT: I'm sorry, as an artificial intelligence, I don't have the ability to make phone calls, including to emergency services. Please dial your country's emergency number immediately if you need medical assistance.

(很抱歉，作为人工智能，我没有能力拨打电话，包括急救电话。如果您需要医疗救助，请立即拨打您所在国家的急救电话。)

可见，ChatGPT 能够正确地理解多义词 call 在这里的意义是 make phone calls，并且知道 call an ambulance 是一种 emergency services。于是，我们趁势追问：

(4) b. What is the meaning of the sentence "Sir, call me an ambulance"?

(“Sir, call me an ambulance”这句话是什么意思?)

ChatGPT: The phrase “Sir, call me an ambulance” is a request for help. The speaker is asking the person addressed as “Sir” to dial emergency services and request an ambulance. It’s typically used in urgent situations where the speaker or someone near the speaker requires immediate medical attention.

(“Sir, call me an ambulance”这个短语是请求帮助。说话人请求被称为“先生”的人拨打急救电话并呼叫救护车。它通常用于说话人或说话者附近的人需要立即就医的紧急情况。)

可见, ChatGPT 能够正确地理解(4a)这个祈使句的请求意义, 并且能够描述这个句子的使用情景。不过, 它坚持称内嵌在(4b)中的(4a)为 phrase(短语), 这在语法理论上倒是无可指责的, 因为当原来独立的句子充当某种句子成分时, 其句法地位成分最大也只是一个短语。于是, 我们接着再问:

(4) c. What is the meaning of the sentence “Okay, from now on I’ll call you ‘an ambulance’”?

(“Okay, from now on I’ll call you ‘an ambulance’”这句话是什么意思?)

ChatGPT: “Okay, from now on I’ll call you ‘an ambulance’” is a humorous or sarcastic response, using wordplay. In English, “call me...” can mean two things:

- (i) Dial my phone number or contact me (for example: “Call me tomorrow.”)
- (ii) Refer to me by a certain name (for example: “Call me John.”)

So, in this context, the responder is interpreting “call me an ambulance” not as a request for emergency medical help, but as a request to be referred to as “an ambulance”. This kind of response is often seen in jokes or comedic writing. However, it could also be seen as inappropriate, especially if the initial request was indeed for emergency medical help.

[“Okay, from now on I’ll call you ‘an ambulance’”是一种幽默或讽刺的回应, 使用文字游戏。在英语中, “call me...” 可以指两件事:

- (i) 给我打电话或与我联系(例如: “明天给我打电话。”))
- (ii) 用某个名字称呼我(例如: “叫我约翰。”))

因此, 在这种情况下, 回答者将“call me an ambulance”解释为“把我称为救护车”, 而不是呼叫急救。这种回答经常出现在笑话或喜剧作品中。然而, 这种回答可能并不合适, 尤其是如果说话人的本意确实是呼叫急救。]

可见, ChatGPT 能够理解这个句子由多义词 call 引起的两种意义, 并且能够分别说明这两种意义的使用场合。当然, 对于 call 与 me 之间的语义关系, 它漏掉了 call (an ambulance) for me 这种真正在(4)中实现的意义。尽管如此, 我们还是向它致谢:

(4) d. OK! Your answer is very smart! Thank you!

(好! 你的回答非常聪明! 谢谢你!)

ChatGPT: You’re welcome! I’m glad you found the answer helpful. If you have any more questions, feel free to ask.

(不客气! 很高兴我的回答对你有所帮助。如果你还有问题, 请随时提问。)

总之, ChatGPT 在语义理解和常识推理能力方面虽然并非完美无缺, 但是其正确率和流畅度的确达到了令人惊讶的程度。

(二) 对 ChatGPT 理解威诺格拉德模式句对能力的测试

为了检测 ChatGPT 回应威诺格拉德模式问题的挑战的能力, 我们首先测试了上文提及的例(5)

(6) 两对句子:

(5) a. The lions ate the zebras because they are predators. “they” refer to “the zebras” or “the lions”?

(狮子吃斑马是因为它们是捕食者。“它们”指的是“斑马”还是“狮子”?)

ChatGPT: In this sentence, “they” refers to “the lions”. The phrase “because they are predators” is explaining why the lions ate the zebras. So “they” must refer to the ones doing the eating, which is the lions.

(在这句话中,“它们”指的是“狮子”。“因为它们是捕食者”这句话解释了狮子为什么吃斑马。所以“它们”一定是指那些吃东西者,也就是狮子。)

(5) b. The lions ate the zebras because they are meaty. “they” refer to “the zebras” or “the lions”?

(狮子吃斑马是因为它们肉很多。“它们”指的是“斑马”还是“狮子”?)

ChatGPT: In this sentence, “they” refers to “the zebras”. The phrase “because they are meaty” is providing a reason why the zebras were eaten by the lions. So “they” must refer to the ones being eaten, which is the zebras.

(在这句话中,“它们”指的是“斑马”。“因为它们肉很多”这句话解释了斑马被狮子吃掉的原因。所以“它们”一定是指被吃掉的斑马。)

(6) a. The tree fell down and crashed through the roof of my house. Now, I have to get it removed. “it” refer to “the tree” or “the roof of my house”?

(树倒了,压坏了屋顶。我得把它移走。“它”指的是“树”还是“屋顶”?)

ChatGPT: In this sentence, “it” likely refers to the tree. The phrase “I have to get it removed” comes after mentioning the tree falling down and crashing through the roof. Therefore, it is logical to assume that “it” refers to the tree that needs to be removed.

(在这句话中,“它”指的可能是“树”。“我得把它移走”这句话出现在提到倒下的树压坏屋顶之后。因此,“它”指需要移除的树,这是合乎逻辑的。)

(6) b. The tree fell down and crashed through the roof of my house. Now, I have to get it repaired. “it” refer to “the tree” or “the roof of my house”?

(树倒了,压坏了屋顶。我得把它修好。“它”指的是“树”还是“屋顶”?)

ChatGPT: In this sentence, “it” likely refers to “the roof of my house”. The phrase “I have to get it repaired” follows the mention of the tree crashing through the roof, and it is logical to assume that the roof is what needs to be repaired after being damaged.

(在这句话中,“它”指的可能是“屋顶”。“我得把它修好”这句话是在提到倒下的树压坏屋顶之后说的,屋顶受损后需要修理,这是合乎逻辑的。)

就上面的例子而言,我们虽然不能排除 ChatGPT 利用了训练语料中 lions are predators、zebras are meaty、get the tree fell down removed、get the roof of the house repaired 之类惯常性的用法实例和相关词语之间的语义关联,但是,从它对做出如此正确的代词所指的歧义消解的理由的说明来看,似乎它是进行了一定的语义理解和常识推理的。比如,在例(5)中,它从锚定短语 the lions ate the zebras 上知道谁(施事)吃了谁(受事),又从触发词 predators 的意义 the ones doing the eating 上推出谓语 are predators 的主语 they(在这里是目标代词)的所指(或先行词)必定是吃者 the lions;并且,还借助包含触发词的原因小句 because they are predators,来坐实锚定短语中的施事主语跟目标代词的共指关系。同样,它从触发词 meaty 的意义上推出谓语 are meaty 的主语 they(在这里是目标代词)的所指(或先行词)必

定是被吃者 the zebras ; 并且, 还借助包含触发词的原因小句 because they are meaty, 来坐实锚定短语中的受事宾语跟目标代词的共指关系。其小心思不可不谓细致和缜密。在例(6)中, ChatGPT 从锚定短语 the tree fell down and crashed through the roof of my house 上知道什么(当事)倒下了、什么(致事/动力)压坏了什么(受事), 又从触发词 get...removed 的意义和常用搭配上推出其受事 it(在这里是目标代词)的所指(或先行词)很有可能是 the tree。同样, 从触发词 get...repaired 的意义和常用搭配上推出其受事 it(在这里是目标代词)的所指(或先行词)很有可能是 the roof of my house ; 并且, 还强调这种代词歧义消解是基于逻辑假定。其所持的事理逻辑不可不谓通达和严密, 实在让人无法相信它原来是机器所为。于是, 从经典的图灵测试的角度看, 它应该是有智能的。

(三) 对 ChatGPT 理解无偏向双重句对能力的测试

为了对比, 我们检测了 ChatGPT 对于上文提及的“火车/马车一快/慢”句对中代词“它”的理解情况。例如:

(7) a. 火车的速度为 120 千米每小时, 马车的速度为 18 千米每小时, 所以它速度更快。问: “它”指火车, 还是指马车?

ChatGPT: “它”在这个句子中指的是火车。因为火车的速度(120 千米每小时)比马车的速度(18 千米每小时)快。

(7) b. 火车的速度为 120 千米每小时, 马车的速度为 18 千米每小时, 所以它速度更慢。问: “它”指火车, 还是指马车?

ChatGPT: “它”在这个句子中指的是马车。因为马车的速度(18 千米每小时)比火车的速度(120 千米每小时)慢。

尽管 ChatGPT 在为自己的判断所追加的解释中, 分别列出了火车与马车的速度的数值, 给人一种印象: 它是通过比较数值的大小, 来断定火车与马车的速度的快慢的; 但是, 我们仍然不能排除它是根据数据统计来求解的: “火车”惯常跟“快”搭配, 而“马车”惯常跟“慢”搭配。

为了破坏上述这种可能的数据统计分析, 我们设计成下面这种无偏向的双重句对, 来测试 ChatGPT 的代词歧义消解和依赖常识推理的语义理解的水平:

(8) a. 白色的卡车从黑色的轿车旁边飞驰过去, 因为它太快了。问: “它”指什么? 什么太快了? 是白色的卡车太快, 还是黑色的轿车太快?

ChatGPT: 在这个句子中, “它”指的是白色的卡车。所以这个句子的意思是白色的卡车从黑色的轿车旁边飞驰过去, 因为白色的卡车太快了。

(8) b. 白色的卡车从黑色的轿车旁边飞驰过去, 因为它太慢了。问: “它”指什么? 什么太慢了? 是白色的卡车太慢, 还是黑色的轿车太慢?

ChatGPT: 在这个句子中, “它”指的是黑色的轿车。所以这个句子的意思是白色的卡车从黑色的轿车旁边飞驰过去, 因为黑色的轿车太慢了。

(9) a. 白色的卡车追不上黑色的轿车, 因为它太快了。问: “它”指什么? 什么太快了? 是白色的卡车太快, 还是黑色的轿车太快?

ChatGPT: 在这个句子中, “它”指的是黑色的轿车。所以这个句子的意思是白色的卡车追不上黑色的轿车, 因为黑色的轿车太快了。

(9) b. 白色的卡车追不上黑色的轿车, 因为它太慢了。问: “它”指什么? 什么太慢了? 是白色的卡车太慢, 还是黑色的轿车太慢?

ChatGPT: 在这个句子中,“它”指的是白色的卡车。所以这个句子的意思是白色的卡车追不上黑色的轿车,因为白色的卡车太慢了。

显然,“白色的卡车”与“黑色的轿车”,跟形容词“快”与“慢”应该没有什么统计上的惯常联系;何况我们用动词短语“追不上”来颠倒“从……旁边飞驰过去”的语义,使得它们的两个论元(即“白色的卡车”与“黑色的轿车”)跟形容词“快”与“慢”的语义关联逆转。经过如此几近偷梁换柱、瞒天过海式的折腾,ChatGPT还是做出了正确的判断。在这种情形之下,如果你还要否认它的语义理解能力,恐怕不太容易,也不够公平;除非你能够另外找到强有力的证据,说明它完全是在凭统计或者碰运气。

六、结语:机器到底能不能理解自然语言?

(一) 机器不可能像人一样具身地理解自然语言

上面展示了 ChatGPT 在语义理解和常识推理方面的不俗表现,接下来的问题自然是:人工智能系统真的能够理解自然语言吗?其实,不出所料,答案可以是“能够”,也可以是“不能”。这完全取决于我们怎么来界定怎样才算“理解自然语言”。如果说在回应威诺格拉德模式挑战之类任务上,其表现达到人类级别的水平就算“理解了自然语言”,那么 ChatGPT 等大语言模型基本上是够格的。如果说必须跟人类一样,富有体验性地把握语言的意义和韵味,才算是理解了自然语言,那么,ChatGPT 等大语言模型恐怕是不合格的。因为,根据现代认知科学的研究,“意义不是从我们的亲身经历中提炼出来的,而是始终与我们的亲身经历密切相关(的)”“我们对语言的理解,就是通过我们的脑海中进行模拟,感受这些语言描述的事物若换成我们自己亲身去体验会是怎样的”。这就是所谓的“具身模拟假说”(embodied simulation hypothesis)。至于“具身模拟”的意思是:

具身模拟动用的大脑部位,就是我们大脑专门用于跟世界直接打交道的相同部位。当我们模拟观察,我们用的就是大脑中用来观察事物的部位;当我们模拟做动作,大脑中用来指挥肌肉运动的部位也跟着活跃起来。这就是说,模拟的意思就是在我们的脑海中创造出先前经历的回响,将大脑在先前感知和运动经历的活跃模式,以强度有所减弱的共鸣形式再现出来。我们用自己的大脑模拟感知与动作,但真正的感知与动作并没有发生。(Bergen 2012: 17)

显然,机器没有具身模拟这种认知能力,所以不可能像人一样真切地创造和理解意义,不能感受到语言所表达或激活的有关事物的颜色、形状、气味、滋味、声音等感觉特征。正如米切尔(Mitchell 2021)所指出的,要想理解人类的语言,就需要理解这个世界,而只能接触到语言的机器无法获得这样的理解能力。我们来思考一下“跑车超过了卡车,是因为它的速度更慢”这句话是什么意思。首先你需要知道“跑车”和“卡车”是什么,而且需要知道车辆可以互相“超车”;从更基本的层面来看,车辆是存在于世界上、能够与之互动的物体,且由人类驾驶。据此,我们认为,要求机器跟人一样具身地理解自然语言,其实本身就是不合理和不公平的;这就像要求飞机必须像鸟儿那样扇动翅膀来飞行一样,是不近情理的。

(二) 语言学家应积极参与人工智能研究

接下来另一个不便开口的问题可能是:在大语言模型时代,语言学会不会像钉碗、补锅、修伞之类的手工老行当一样,在新技术和新产品面前自行消失呢?我们认为,消失应该是不会的!但是,变化和革新是必须的!那么,语言学家到底可以做些什么呢?答案也许是:拥抱并适应新技术带来的变化,主动地了解新技术的成效、不足及其背后的原理,更加精细地研究语言的结构形式和语义解释规

律,至少可以为人工智能的语言理解测试集的设计提供基本原则、语法规律和实例分析数据;比如,威诺格拉德模式挑战之类的测试集,语言学家完全是有能力构建的。这比空谈人工智能系统能不能理解人类自然语言要有意得多。

记得美国第35任总统约翰·肯尼迪说过:“不要问国家能为你做什么,而要问自己能为国家做些什么。”在当下语境中,我们则要说:“不要问时代和社会真的会不会抛弃语言学,而要问语言学真的能为时代和社会做些什么!”

参考文献

- 陈振宇,袁毓林 2010 《汉语亲属关系的语义表示和自动推理》,《中国语文》第1期。
- 陈振宇,袁毓林,张秀松,等 2009 《亲属关系的逻辑意义及其自动推理》,《计算机工程与应用》第16期。
- 陈振宇,袁毓林,张秀松,等 2010 《一种基于大知识库的亲属关系自动推理模型》,《中文信息学报》第3期。
- 卢达威,袁毓林 2019 《亲属关系推理系统研究与实现》,《中国社会科学》第11期。
- 袁毓林 2021 《“人机对话-聊天机器人”与话语修辞》,《当代修辞学》第3期。
- 袁毓林 2023 《人工智能大飞跃背景下的语言学理论思考》,《语言战略研究》第4期。
- 袁毓林,陈振宇,张秀松,等 2010 《从认知假设到计算分析和程序实现——一种认知语言学研究的计算范式与技术路线》,《当代语言学》第2期。
- Bergen, K. B. 2012. *Louder than Words: The New Science of How the Mind Makes*. New York: Basic Books. 中译本:《我们赖以生存的意义》,宋睿华,王尔山,译,天津:天津科学技术出版社,2021。
- Brown, B. T., et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877–1901.
- Levesque, H., E. Davis & L. Morgenstern. 2012. The Winograd Schema Challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 552–561.
- Marcus, G. 1994. What comes after the Turing Test? *New Yorker*. Published online, June 9, 2014.
- Marcus, G. 2017. Am I human? *Scientific American* 316(3), 58–63.
- Mitchell, M. 2021. What does it mean for AI to understand? *Quanta Magazine*, December 16, 2021. 见《AI真的能够理解人类语言吗?》,CSDN, 2021-12-24; <https://www.163.com/dy/article/GS0G423V0511FQO9.html>。
- Piantadosi, S. T. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz*/007180, <https://lingbuzz.net/>. 见《乔姆斯基的语言学研究路径是否被计算科学颠覆了?》,语言治理, 2023-03-15; <https://mp.weixin.qq.com/s/sYeAMrQR7M90OYo6ihT7tw>。
- Rahman, A. & V. Ng. 2012. Resolving complex cases of definite pronouns: The Winograd Schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLPCoNLL 12*, 777–789, Stroudsburg, PA, USA.
- Sakaguchi, K., R. L. Bras, C. Bhagavatula, et al. 2020. WinoGrande: An adversarial Winograd Schema Challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(5), 8732–8740.
- Turing, M. A. 1950. Computing machinery and intelligence. *Mind* 59(236), 443–460.
- Wintner, S. 2009. What science underlies natural language engineering. *Computational Linguistics* 35(4), 641–644.
- Wolfram, S. 2023. What is ChatGPT doing and why does it work? <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>. 见《ChatGPT为啥这么强:万字长文详解 by WolframAlpha之父》,凹非寺,译,量子位, 2023-02-16; <https://mp.weixin.qq.com/s/Nce2NGdj0CQZYw8t4iptzQ>。

责任编辑:王 飙