

语义理解与常识推理的机器表现 和人类基线之比较^{*}

——怎样评估 ChatGPT 等大型语言模型的语言运用能力？

袁毓林

提 要 本文讨论怎样通过跟人类基线的比较,来合理地评估 ChatGPT 等现代大型语言模型的语言运用能力。首先,用代词指称歧义和否定辖域问题测试 ChatGPT,展示语言大模型在语义理解和常识推理方面的优秀表现;接着简介维诺格拉德模式挑战及其升级版 WinoGrande 数据集,还介绍了我们对于这种类型的测试题和机器表现的评估方式的改进方案(把仅触发词不同的“句对”扩展为锚定词也不同的“句偶”,把机器表现跟人类被试的表现进行比较);然后介绍我们怎样用“句对”和“句偶”测试 ChatGPT 和人类被试,并且把人类和机器的表现进行对比,从而得出语言大模型的语言运用能力接近人类的结论。

关键词 语义理解/常识推理 ChatGPT/大型语言模型 维诺格拉德模式/句对与句偶 机器表现/人类基线

○ 引言:怎样衡量语言大模型的语言运用能力?

凡是用过 ChatGPT 等现代大型语言模型(large language model,LLM)的人,可能都会惊叹于其在语言理解与生成方面的强大功能——聊天、答问、写作,甚至逻辑推理、解数学题、编程序、写代码;更何况,它们是以目前并不为人们(甚至人工智能专家)所完全理解的方式和原理进行工作的(详见 Lee and Trott 2023),并且还几乎绕开了当代主流的语言学理论与方法(详见 Piantadosi 2023)。那么,ChatGPT 等大模型的语言运用能力到底有多强?能不能用一种在语言学上合理的方式来测试和评估?这也许是许多语言学工作者所普遍关心的问题。

本文尝试根据笔者有限的文献阅读和粗浅的实验调查,来讨论和回答这些问题。为了方便,我们将主要关注语言大模型的语义理解和常识推理能力的测试集的建构和实施问题。文章的第一节,通过用一组跟代词指称歧义的消解和否定辖域的确定等相关的问题来测试 ChatGPT,从而展示现代大型语言模型在语义理解和常识推理方面的优秀表现。第二节简介“新图灵测试”中跟语言学相关的维诺格拉德模式挑战(Winograd Schema Challenge, WSC),

作者简介:袁毓林,男,江苏昆山人,博士,教授,研究方向为理论语言学和汉语语言学,特别是句法学、语义学、语用学、计算语言学和中文信息处理,电邮: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(编号:CPG2024-00005-FAH)和启动研究基金(编号:SRG2022-00011-FAH)的资助,还承博士生丛冠良同学帮助调查,谨此一并致以谢忱。

其宗旨在于测试人工智能系统的语义理解和常识推理的能力;接着简介 WSC 的进一步升级与扩展的版本,即 WinoGrande 数据集;还介绍了我们对于测试语言大模型的试题的改进方案,即从原先只有触发词不同的两个句子组成的“句对”上,改编和扩展出锚定词也不同的四个句子组成的“句偶”,从而增加句子的语义困惑度(perplexity),更加有效地排除机器利用词语之间的选择限制或统计关联的可能性。第三节介绍我们为了检测 ChatGPT 回应维诺格拉德模式挑战的能力,怎样挑选和改编了一些英语和汉语的句子,形成意义复杂和微妙的“句对”和“句偶”来对它进行测试;详细介绍我们怎样用这些句子对 68 位大学生和研究生进行测试,从而形成人类基线(human baseline),来比较客观地衡量 ChatGPT 的语义理解和常识推理能力。第四节是一个简短的总结,在上文比较机器表现与人类表现的基础上,得出结论:在纯粹的语言理解方面,机器并不输于人类;但是,在需要利用常识推理的语言理解方面,机器是不如人类的。因此,折中一下,也许可以大概地说:语言大模型的语言运用能力接近于人类水平。

一 ChatGPT 的语义理解与常识推理水平有多高?

1.1 虽然 ChatGPT 是一款专注于语言生成和聊天交谈的人工智能系统,但是它在语义理解和常识推理方面的表现也不容小觑。比如,人类自然语言中代词的指称歧义的消解及其条件问题(pronoun disambiguation problem, PDP),历来是句法学和语义学的核心话题,也是检验各种语法学和语义学理论的试金石。(详见徐烈炯 1995:227—253)然而,ChatGPT 似乎可以轻而易举地识别有关代词的先行语,得出其正确的指称意义解读,并且还能够说明做出这种判断的理由。例如:

(1)a. 老刘给小刘打电话,但是他没有打通。

问题:他指谁?/谁没有打通电话?/是老刘没有打通电话,还是小刘没有打通电话?

ChatGPT:在这个句子中,“他”很明显地指的是老刘。所以这个句子的意思是老刘给小刘打电话,但是老刘没有打通。

b. 老刘给小刘发短信,但是他没有回应。

问题:他指谁?/谁没有回应?/是老刘没有回应短信,还是小刘没有回应短信?

ChatGPT:在这个句子中,“他”很明显地指的是小刘。所以这个句子的意思是老刘给小刘发短信,但是小刘没有回应。

在例(1)的两个句子中,纯粹从句法结构上看,转折分句中的代词“他”既可回指条件分句的主语“老刘”,又可回指条件分句中的介词宾语“小刘”;但是,ChatGPT 能够根据语义关系来确定:“没有打通(电话)”的主体(theme)就是“打电话”的施事(agent),即“老刘”;而“没有回应(短信)”的主体却是“(老刘)(给他)发短信”的与事(dative)或目标(goal),即“小刘”。再如:

(2)a. 我把水从瓶子里倒进杯子里,直到它满了。

问题:什么是满的,瓶子还是杯子?

ChatGPT:在这个句子中,“它”指的是杯子。因为你正在把水从瓶子倒入杯子,所以当你说“直到它满了”,你是在描述你停止倒水的条件,也就是杯子被装满。

b. 我把水从瓶子里倒进杯子里,直到它空了为止。

问题:什么是空的,瓶子还是杯子?

ChatGPT:在这个句子中,“它”指的是瓶子。因为你正在把水从瓶子倒入杯子,所以当你说“直到它空了为止”,你是在描述你停止倒水的条件,也就是瓶子被倒空。

可见,ChatGPT 不仅能够消解代词“它”的指称歧义,而且还能够说明为什么这样做的理由。其中,除了利用一定的句法语义知识之外,还借助了常识的支持。再如:

(3)a. 张三的叔叔在网球上仍能打败他,尽管他比他大 30 岁。

问题:谁更老?/谁更老,张三还是张三的叔叔?

ChatGPT:在这个句子中,“他比他大 30 岁”这部分指的是张三的叔叔比张三年长 30 岁。所以,张三的叔叔比张三更老。

b. 张三的叔叔仍然可以在网球上击败他,尽管他比他年轻 30 岁。

问题:谁更年轻?/谁更年轻,张三还是张三的叔叔?

ChatGPT:在这个句子中,“他比他年轻 30 岁”这部分指的是张三比他的叔叔年轻 30 岁。所以,张三比他的叔叔更年轻。

可见,ChatGPT 能够识别这两个后置的转折分句中代词“他”的指称歧义,而且还能够说明这样做是利用了形容词“大(年长)”和“年轻”的意义。另外,我们估计它还利用了“叔叔”的语义知识来推定“大(年长)”的是“张三的叔叔”,而“年轻”的则是“张三”。

1.2 除了代词指称的歧义消解问题,否定的辖域及其语义解释也是句法学和语义学的核心问题。同样,ChatGPT 在理解否定表达方面的优秀表现也令人惊讶。例如:

(4)a. 这一阵子,他爸爸天天不上班。

问题:他爸爸这一阵子上过班,真的还是假的?/他爸爸上班的事件,发生过没有?

ChatGPT:根据句子 a“这一阵子,他爸爸天天不上班”,我们可以理解为在这段时间内,他的父亲一直都没有去上班。所以,我们可以说他爸爸这一阵子上过班的事件并没有发生,这是假的。

b. 这一阵子,他爸爸不天天上班。

问题:他爸爸这一阵子上过班,真的还是假的?/他爸爸上班的事件,发生过没有?

ChatGPT:根据句子 b“这一阵子,他爸爸不天天上班”,我们可以理解为在这段时间内,他的父亲并没有每天都去上班,但这并不排除他有些天还是上班的。所以,我们可以说他爸爸这一阵子上过班的事件是发生过的,这是真的。

可见,ChatGPT 能够根据否定副词“不”与时间副词“天天”的顺序来确定否定的辖域(范围),从而获得句子的语义解释“他爸爸上班的事件发生过/没有发生”;并且,据此来判断问题中提出的命题“他爸爸这一阵子上过班”的真/假。

为了增加语义理解的难度,我们从文献上寻找了一对需要百科知识才能理解其意义的否定句,并进行改编来测试 ChatGPT:

(5)a. 如果一只猫的体温低于平均水平,它就不在危险之中。

问题:这是真的还是假的?

ChatGPT:这个句子是错误的。猫的正常体温范围在 37.5 至 39.2 摄氏度。如果猫的体温低于这个范围,那可能是有健康问题的标志,比如可能受寒、休克、消化系统问题、感染等等。因此,如果你发现你的猫体温低于正常范围,应立即联系兽医以确定可能的原因并获取合适的治疗。

b. 如果一只猫的体温低于平均水平,它就不在安全范围。

问题:这是真的还是假的?

ChatGPT:这个句子是真的。猫的正常体温范围在 37.5 至 39.2 摄氏度。如果一只

猫的体温低于这个范围,那可能意味着它有健康问题,比如可能受寒、休克、消化系统问题、感染等等。因此,如果你发现你的猫体温低于正常范围,应立即联系兽医以确定可能的原因并获取合适的治疗。

可见,ChatGPT 不仅能够理解“危险”和“安全”这对反义词的意义,理解否定表达“不在危险之中”和“不在安全范围”的意义相反;而且,还拥有关于猫的体温及其健康情况的百科知识。为了检验它是不是真的具有上述语言知识,我们把例(5)改编成下面的肯定句来测试:

(6)a. 如果一只猫的体温低于平均水平,它就处在危险之中。

问题:这是真的还是假的?

ChatGPT:这个句子是真的。猫的正常体温范围在 37.5 至 39.2 摄氏度。如果一只猫的体温低于这个范围,那可能意味着它有健康问题,可能包括受寒、休克、消化系统问题、感染等等。因此,如果你发现你的猫体温低于正常范围,应立即联系兽医以确定可能的原因并获取合适的治疗。

b. 如果一只猫的体温低于平均水平,它就处在安全范围。

问题:这是真的还是假的?

ChatGPT:这个句子是错误的。猫的正常体温范围在 37.5 至 39.2 摄氏度。如果一只猫的体温低于这个范围,那可能意味着它有健康问题,可能包括受寒、休克、消化系统问题、感染等等。因此,如果你发现你的猫体温低于正常范围,应立即联系兽医以确定可能的原因并获取合适的治疗。

可见,对于反义形容词及其否定表达,对于否定句及其相应的肯定句,ChatGPT 都能进行正确的语义解读;并且,它能够利用有关的百科知识来判断相关句子的真假。

二 作为“新图灵测试”的“维诺格拉德模式”及其改进方案

2.1 在上一节中,我们用例(1)–(6)这种近于最小差别对(minimal pair)的两个语句,来测试 ChatGPT 的语义理解和常识推理的水平。这种类似双胞胎的两个句子(twin sentences),通常只相差一两个词语,但是所指的人、事物或句子的真假就完全不同。这种句子对(可以简称为“句对”),在人工智能社区被称为维诺格拉德模式(Winograd Schema, WS)。因为,著名的人工智能前辈 Terry Winograd 在 1971 年提出了这种模式;在相同的场景下提出不同的问题,问句中的代词存在着指称歧义(在句法上可以回指场景描述小句中不同的先行语,但是在语义上只能回指其中的一个先行语)。例如(见 Marcus 2017):

(7) a. The city councilmen refused the demonstrators a permit because they feared violence.

Question: Who feared violence?

b. The city councilmen refused the demonstrators a permit because they advocated violence.

Question: Who advocated violence?

我们把例(7)改写为中文版本,用以对 ChatGPT 进行测试,结果如下:

(8)a. 市政府拒绝给示威者发放游行许可证,因为他们担心暴力冲突。

问题:他们指谁?/谁担心暴力冲突?/是市政府担心暴力冲突,还是示威者担心暴力冲突?

ChatGPT:在这个句子中,“他们”很明显地指的是市政府。所以这个句子的意思是市政府拒绝给示威者发放游行许可证,因为市政府担心会发生暴力冲突。

b. 市政府拒绝给示威者发放游行许可证,因为他们鼓动暴力冲突。

问题:他们指谁?/谁鼓动暴力冲突?/是市政府鼓动暴力冲突,还是示威者鼓动暴力冲突?

ChatGPT:在这个句子中,“他们”很明显地指的是示威者。所以这个句子的意思是市政府拒绝给示威者发放游行许可证,因为示威者鼓动暴力冲突。

可见,ChatGPT能够消解代词“他们”的歧义,顺利地通过维诺格拉德模式测试。

从上面的例子可见,维诺格拉德模式的句对设计十分巧妙,其形式结构特征可以从整体句型、先行小句和后续小句三个方面来刻画:1)为了差别的最小化,句对中的两个句子采用相同的句型(或结构模式);为了便于检测人工智能系统的常识推理能力,往往都采用由条件小句和结果小句构成的偏正复句;并且,这种复句允许“条件—结果”或“结果—原因”两种相反的语序形式;2)在前面的先行小句(上文)通常是一个句对中两个句子的共有成分,往往是一种表示共同场景的背景性成分,可以称为锚定性小句(anchor clause);并且,其中一定包含后续小句(下文)中的代词可以回指的两个(或多个)可能的先行语;比如,上例的“市政府拒绝给示威者提供游行许可”是锚定性小句,其中的关键词“拒绝”是锚定词(anchor words),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;并且,先行小句中的两个名词“市政府(人士)”和“示威者”,在句法上都可以充当后续小句中的代词“他们”的先行语;3)在一个句对的两个后续小句中,都有一个需要歧义消解的目标代词(target pronoun),它可以回指前面的锚定小句中两个(或多个)先行语中的任意一个;此外,还各有一个限定目标代词的真正所指的触发词语(trigger words),比如上例中的“担心(暴力冲突)”或“鼓动(暴力冲突)”,它们可以帮助人们或人工智能系统对目标代词进行歧义消解,确定其在语义和事理上正确的先行语。正是通过这种复杂又精致的设计,维诺格拉德模式把语言知识和常识推理有机地结合起来,在相当大的程度上可以对人工智能系统的语义理解和常识推理能力进行相对客观的测试。

2.2 值得一提的是,2011年计算机科学家 Hector Levesque、Ernest Davis 和 Leora Morgenstern 用上述维诺格拉德模式,精心设计了 273 个代词歧义消解问题,构成一个测试数据集;并且,提出了维诺格拉德模式挑战(Winograd Schema Challenge, WSC)这种比较客观的测试竞赛。像 1.1 节中的例(1)—(3),都是这个测试集中的经典例子(或者其变体)。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。他们的设计原则是:1)对于人类读者来说,可以轻而易举地消除代词的指称歧义;他们有时甚至未曾觉察到句子有歧义,一下子就确定了代词所回指的先行语;2)消歧无法依靠简单的技巧,比如词语之间的选择限制(如“猎豹”——“捕食者”vs.“羚羊”——“被捕食者”);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(如“火车”——“快”vs.“马车”——“慢”)。^①

实质上,维诺格拉德模式挑战是一种语言理解与常识推理相交叉的测试,可以作为评估机器理解语言所达到的水平的一种有效方式。因此, Marcus (2017) 主张的“新图灵测试”的四种形式中^②,首先选择了维诺格拉德模式挑战,用以测试人工智能系统的语义理解和常识推理的能力。另外,针对现代大型语言模型的测试基准(benchmark) SuperGLUE 竞赛,其中也包含了威诺格拉德模式。尤其值得注意的是,根据 Sakaguchi et al. (2019) 的介绍,他们通过众包流程,构建了由 44,000 个 WSC 风格的问题组成的数据集,形成新的测试基准 WinoGrande,

作为语言大模型迁移学习的资源和测试基准。

令人遗憾的是,语言学家对于语言大模型的语言理解能力的测试工作参与不足。有感于此,我们在学习有关文献并对 ChatGPT 进行测试实践的基础上,对于维诺格拉德模式及其测试提出两点改进与补充:第一,扩充现有的句对,形成两个相应句对构成的“句偶”(a couple of twin sentences),以便更好地排除词语之间的选择偏向或统计关联;第二,同时对人类被试进行测试,形成“人类基线”,以便更好地了解语言大模型的语言理解水平。例如:

(9)a. 保姆根本抱不动那个女孩,因为她太壮实了。

问题:她指谁?/谁太壮实了?/是保姆非常壮实,还是那个女孩非常壮实?

ChatGPT:在句子 a 中,“她”应该指的是那个女孩。那个女孩太壮实了。因为句子中的“她”是跟随在“那个女孩”后面的,所以“她”指的是“那个女孩”,并且句子的意思是因为这个女孩太壮实,保姆无法抱动她。所以,是那个女孩非常壮实,而不是保姆。

b. 保姆根本抱不动那个女孩,因为她太瘦弱了。

问题:她指谁?/谁太瘦弱了?/是保姆非常瘦弱,还是那个女孩非常瘦弱?

ChatGPT:在句子 b 中,“她”应该指的是保姆。保姆太瘦弱了。句子的意思是保姆因为太瘦弱,无法抱动那个女孩。所以,是保姆非常瘦弱,而不是那个女孩。

可见,ChatGPT 能够消解代词“她”的指称歧义,正确地找到其先行语;但是,给出的理由不能成立。因为,例(9a)—(9b)中“她”都跟在名词短语“那个女孩”后面,但指称是不同的。这种结论正确而理由偏差,可能源于训练语料中有不少诸如此类的外行的语法分析材料。

传统的维诺格拉德模式,主要是通过改变后续小句中的触发词语(如把例 8a 中的“壮实”改变为“瘦弱”)来构建句对的。但是,我们尝试同时改变先行小句中的锚定词语来构建句对;比如,还把例(9)中的“根本抱不动”改变为“一把就抱起了”,来构建相应的新的句对:

(10)a. 保姆一把就抱起了那个女孩,因为她太壮实了。

问题:她指谁?/谁太壮实了?/是保姆非常壮实,还是那个女孩非常壮实?

ChatGPT:在句子 a 中,“她”应该指的是保姆。保姆太壮实了。句子的意思是因为保姆非常壮实,所以能够轻易地抱起那个女孩。所以,是保姆非常壮实,而不是那个女孩。

b. 保姆一把就抱起了那个女孩,因为她太瘦弱了。

问题:她指谁?/谁太瘦弱了?/是保姆非常瘦弱,还是那个女孩非常瘦弱?

ChatGPT:在句子 b 中,“她”应该指的是那个女孩。那个女孩太瘦弱了。句子的意思是因为那个女孩非常瘦弱,所以保姆能轻易地抱起她。所以,是那个女孩非常瘦弱,而不是保姆。

这样,例(9)和(10)两个句对就构成了一个句偶,增加了测试句的语义困惑度,更加有效地排除词语之间的选择偏好或统计关联对语句意义理解的支持。

三 ChatGPT 的语言理解水平与人类基线的比较

3.1 为了检测 ChatGPT 的语言理解能力,我们自拟或从有关文献上挑选和改编了一些英语和汉语的句子^③,形成 80 个问题,向它提问。结果,ChatGPT 回答明显错误的有 4 个,占 5%;对错不太确定的有 4 个,也占 5%。由于 ChatGPT 的这 8 个回答不同于笔者(主试)事先

拟定的标准答案,因而暂时判为错误。这样算下来,机器的正确率约为 90%。当然,这是很不严格的评估。为了更加严格,我们只选择符合维诺格拉德模式的句对和句偶,来形成一套试题,包括 12 道大题、24 道小题、30 个具体问题。为了得到相对客观的比较结果,我们用这套具有一定难度的试题,同时对人类被试进行测试,以便形成人类基线。这些人类被试共 68 人,包括大专生 2 人、本科生 27 人、硕士生 25 人、博士生 12 人、没有填写具体信息的 2 人。^④最终,平均下来,人类被试的正确率约为 85.69%,高于机器的 73.33%。具体的计算方法,详见第 4 节的介绍。下面,我们先来看 ChatGPT 和人类被试对 2 个英语句对的反应:

(11)a. Ann asked Mary what time the library closes, because she had forgotten. She refer to Ann or Mary?

ChatGPT: In this sentence, “she” likely refers to Ann. The phrase “because she had forgotten” is giving a reason for Ann’s action (asking Mary what time the library closes), so it makes sense that “she” refers to the person who is doing the asking, which is Ann.

b. Ann asked Mary what time the library closes, but she had forgotten. She refer to Ann or Mary?

ChatGPT: In this sentence, “she” could refer to either Ann or Mary, as it’s not clear from the sentence alone. If “she” refers to Ann, it could mean that Ann asked the question but then realized that she had forgotten the answer. If “she” refers to Mary, it could mean that Ann asked Mary the question, but Mary had forgotten the answer. Without more context, it’s not possible to definitively determine who “she” refers to. [袁案:半错,因为它交代了两种解读的理由,以下同此]

对于例(11a),机器的表现(100%正确)稍优于人类被试;虽然有 67 人选择 Ann,但有 1 人选择 Mary,正误比为 98.53% : 1.47%。对于例(11b),机器的表现(两可)远逊于人类被试;有 56 人选择 Mary,只有 12 人选择 Ann,正误比为 82.36% : 17.65%。这说明,机器不能正确理解连词“but”表示转折关系,所以无法借此进行有关的推理:Ann 问 Mary 图书馆闭馆的时间,Ann 预期 Mary 知道图书馆闭馆的时间;但是,出乎 Ann 意料的是 Mary 居然忘了图书馆闭馆的时间。至少在这一题上,人类的面子还是挣回来了不少。

(12)a. Once in Poland, Dennis enjoyed the trip more than Jason, because he had a shallow understanding of the Polish language. Ask: who is he?

ChatGPT: In this sentence, “he” most likely refers to Dennis, because the pronoun “he” is typically used to refer back to the most recently mentioned person. Therefore, Dennis is the one who had a shallow understanding of the Polish language.

b. Once in Poland, Dennis enjoyed the trip more than Jason, because he had a deeper understanding of the Polish language. Ask: who is he?

ChatGPT: In this sentence, “he” is more likely to refer to Dennis. The phrase “Dennis enjoyed the trip more than Jason” implies that there’s something that made Dennis enjoy the trip more. Given that the following clause is “because he

had a deeper understanding of the Polish language”, it’s reasonable to infer that “he” refers to Dennis, whose enjoyment was greater due to his understanding of the language.

对于例(12a),机器的表现跟人类被试持平,都是不确定;有35人选择Dennis,有33人选择Jason,正误比为51.47%:48.53%。但是,机器给出的理由“the pronoun ‘he’ is typically used to refer back to the most recently mentioned person”显然错误,因为例(12a)中离代词“he”最近的刚刚提及的人是Jason。这也暴露了语言模型在训练时不求甚解的弱点。对于例(12b),机器的表现(100%正确)略胜于人类被试;有64人选择Dennis,只有4人选择Jason,正误比为94.12%:5.86%。这说明,人有时因为疲劳等原因会犯迷糊,而机器一般不会。

3.2 接下来,我们再来看ChatGPT和人类被试对5个汉语句对的反应:

(13)a.那个圆球碰到那张桌子以后就碎了,因为它用泡沫塑料做的。

问题:它指什么?/什么碎了?/是那个圆球还是那张桌子碎了?/什么用泡沫塑料做的?/是那个圆球还是那张桌子,是用泡沫塑料做的?

ChatGPT:在句子a中,“它”应该是指那个圆球。那个圆球碎了。句子的意思是那个圆球碰到那张桌子后就碎了,因为它是由泡沫塑料制成的。所以,是那个圆球是用泡沫塑料做的,而不是那张桌子。

b.那个圆球碰到那张桌子以后就碎了,因为它用钢材做的。

问题:它指什么?/什么碎了?/是那个圆球还是那张桌子碎了?/什么用钢材做的?/是那个圆球还是那张桌子,是用钢材做的?

ChatGPT:在句子b中,“它”的指代对象可能有一些含糊,但根据常规理解和语境,“它”应该是指那张桌子。那个圆球碎了。句子的意思是那个圆球碰到了那张桌子之后就碎了,因为桌子是由钢材制成的,硬度大。所以,是那张桌子是用钢材做的,而不是那个圆球。

对于例(13a)中的两个关键问题,机器的回答(100%正确)比人类被试好一些;有61人认为“它”指“圆球”,有7人认为“它”指“桌子”(89.71%:10.29%);有63人认为“碎了”的是“圆球”,有5人认为“碎了”的是“桌子”(92.65%:7.35%)。对于例(13b),机器的表现虽然有些迟疑,但是最终的回答(100%正确)比人类被试好得多;人类被试中有47人认为“它”指“桌子”,有21人认为“它”指“圆球”(69.12%:30.88%);有46人认为“碎了”的是“圆球”,有22人认为“碎了”的是“桌子”(67.65%:32.35%)。这说明,人类有时会被复杂的事理关系给绕糊涂了,而机器反而尽可能按照逻辑进行推导,得出正确的语义解释。再如:

(14)a.律师向证人问了一个问题,但是他拒绝回答。

问题:他指谁?/谁拒绝回答?/是律师拒绝回答,还是证人拒绝回答?

ChatGPT:在这个句子中,“他”很明显地指的是证人。所以这个句子的意思是律师向证人问了一个问题,但是证人拒绝回答。

b.律师向证人问了一个问题,但是他拒绝重复。

问题:他指谁?/谁拒绝重复?/是律师拒绝重复,还是证人拒绝重复?

ChatGPT:在这个句子中,“他”可能有歧义,可能指律师,也可能指证人。从常规语境推断,“他”可能指的是律师,意思是律师拒绝重复问题。然而,也有可能是证人拒绝重复他的回答。要获得准确的理解,需要更多的上下文信息。在写作和交

流中,为了避免这种混淆,应该尽量明确地表达主语,避免使用可能产生歧义的代词。[袁案:半错]

对于例(14a),机器的表现(100%正确)比人类被试好;有65人选择“证人”,有3人选择“律师”(95.59%:4.41%)。对于例(14b),机器的表现(两可,近于盲猜)远逊于人类被试;有52人选择“律师”,有16人选择“证人”(76.47%:23.53%)。这说明,人类会按照事理逻辑进行推导,尽可能常用常规的语义解释;而机器可能按照套路来穷尽代词回指及其语义解释的各种可能性,包括在事理上不成立的可能性。这一点,在下面这个例子上表现得尤为明显:

(15)a. 这款杀毒软件成功检测出新的电脑病毒,因为它太智能了。

问题:它指什么?/它指这款杀毒软件还是新的电脑病毒?/是这款杀毒软件太智能,还是新的电脑病毒太智能?

ChatGPT:在这个句子中,“它”指的是这款杀毒软件。所以这个句子的意思是这款杀毒软件太智能,所以能成功检测出新的电脑病毒。

b. 这款杀毒软件没有检测出新的电脑病毒,因为它太智能了。

问题:它指什么?/它指这款杀毒软件还是新的电脑病毒?/是这款杀毒软件太智能,还是新的电脑病毒太智能?

ChatGPT:在这个句子中,“它”可能有歧义。理论上,“它”可以指这款杀毒软件,也可以指新的电脑病毒。然而,更常见的语境下,“它”应该是指这款杀毒软件。所以,句子可能的含义是这款杀毒软件太过“智能”,以至于它并未能检测出新的电脑病毒。这可能是由于软件的检测算法过于复杂或者是误判了某种行为,导致没有检测出病毒。这是一个猜测,实际含义可能需要更多的上下文信息来确定。

[袁案:半错]

对于例(15a),机器的表现跟人类被试一样好(100%正确)。对于例(15b),机器的表现(两可,近于盲猜)远逊于人类被试;有63人选择“新的电脑病毒”,有5人选择“这款杀毒软件”,正误比为92.65%:7.35%。这说明,人类明白没有被杀毒软件检测出来的电脑病毒是足够智能的这一事理,并且以此进行逻辑推导和语义解释;而机器可能不明白这种事理,只能按照套路来强为之说。

看来,ChatGPT之类大型语言模型在常识推理方面还有很长的道路要走。相对来说,在不怎么依赖常识的纯语义学问题上,它应对得比较从容。例如:

(16)a. 整整一节数学课,小明一直没有开小差。

问题:小明在数学课上开小差了,真的还是假的?/小明开小差的事件,发生了没有?

ChatGPT:根据句子“整整一节数学课,小明一直没有开小差”,我们可以理解为在那节数学课上,小明一直保持专注,没有走神或开小差。所以,我们可以说小明在数学课上开小差的事件并没有发生,这是假的。

b. 整整一节数学课,小明没有一直开小差。

问题:小明在数学课上开小差了,真的还是假的?/小明开小差的事件,发生了没有?

ChatGPT:根据句子“整整一节数学课,小明没有一直开小差”,我们可以理解为在那节数学课上,小明并没有一直开小差,也就是说有一部分时间他可能在开小差,但也有一部分时间他是在专注上课。所以,我们可以说小明在数学课上开小差的

事件是发生了的,这是真的。

可见,对于例(16a),机器知道“开小差”与“走神”意义相近,“没有开小差”与“保持专注”意义相近;对于例(16b),机器知道“没有一直开小差”意味着一部分时间可能在开小差,一部分时间保持专注;并且,能够理解否定词与时间副词的语序跟辖域相关。这也说明了基于分布式语义学的词汇意义的向量表示(包含了词语的位置向量)的成功。^⑤再如:

(17)a. 他们看到了一起车毁人亡的事件。

问题:发生了一起车毁人亡的事件,真的还是假的?

ChatGPT:从提供的句子中,我们无法确定事件的真实性。这个句子只告诉我们“他们看到了一起车毁人亡的事件”,但没有提供其他信息来证明这个事件是真实还是虚假的。如果需要确定事件的真实性,可能需要更多的背景信息或者证据。

b. 他们避免了一起车毁人亡的事件。

问题:发生了一起车毁人亡的事件,真的还是假的?/车毁人亡的事件,发生了没有?

ChatGPT:从提供的“他们避免了一起车毁人亡的事件”这句话看,可以推断,原本有可能发生的车毁人亡事件被他们成功地避免了。所以,严格来讲,没有真正发生一起车毁人亡的事件,只是存在这样一种可能性,但被成功避免了。

c. 如果句子“他们看到了一起车毁人亡的事件”是真的,那么,发生了一起车毁人亡的事件,是真的还是假的?

ChatGPT:如果“他们看到了一起车毁人亡的事件”这句话是真的,那么我们可以理解为确实发生了一起车毁人亡的事件。所以,基于这个前提,我们可以说发生了一起车毁人亡的事件是真的。

对于例(17a),机器一开始并不理解问题的真实意图,所以进行了闪避性的回应;当我们追问了例(17c)以后,它的表现(100%正确)比人类被试略好;有66人选择“真的”,有2人选择“假的”(97.06% : 2.06%)。对于例(17b),机器的表现(100%正确)稍胜于人类被试;有64人选择“假的”,有4人选择“真的”(94.12% : 5.88%)。这说明,对于在语言内部能够通过分布式语义学来解决的问题,机器的语义理解水平相当高,甚至超过人类被试;而对于涉及外部世界的事实和常识推理的问题,机器的表现可能不如人类被试。想一下也是的,这语言大模型就像古代一个背了许多书的秀才,但在人情世故等事理的了解方面,肯定不如我们人类。

3.3 正如前文所说的,为了进一步加大维诺格拉德模式挑战的难度,我们挑选和改编了一些既有的WSC句对;并且,对其中的锚定词或触发词进行反义词替换,构成由两个句对组成的句对骈偶(简称“句偶”);再用这些句偶测试ChatGPT的语义理解和常识推理能力,并且跟人类基线进行对比。下面分别是关于运动方位、打赌规约和情绪表达的句偶:

(18)a. 第一个爬到山顶以后,陈亮把绳子抛下去给了邓忠。

问题:谁第一个爬到山顶?/是陈亮还是邓忠第一个爬到山顶?

ChatGPT:在句子a中,第一个爬到山顶的是陈亮。因为句子的表述是“第一个爬到山顶的陈亮把绳子抛下去给了邓忠”,这意味着陈亮是第一个到达山顶的人。

b. 第一个爬到山顶以后,陈亮把旗子抛上去给了邓忠。

问题:谁第一个爬到山顶?/是陈亮还是邓忠第一个爬到山顶?

ChatGPT:在句子b中,第一个爬到山顶的是邓忠。由于句子的表述是“第一个爬

到山顶以后,陈亮把旗子抛上去给了邓忠”,这意味着邓忠是第一个到达山顶的人,然后陈亮把旗子抛给了他。

(19)a. 第一个爬到山脚以后,陈亮把背包抛下去给了邓忠。

问题:谁第一个爬到山脚?/是陈亮还是邓忠第一个爬到山脚?

ChatGPT:在句子 a 中,第一个爬到山脚的是陈亮。因为句子的表述是“第一个爬到山脚的陈亮把背包抛下去给了邓忠”,这意味着陈亮是第一个到达山脚的人。

[袁案:全错]

b. 第一个爬到山脚以后,陈亮把拐棍抛上去给了邓忠。

问题:谁第一个爬到山脚?/是陈亮还是邓忠第一个爬到山脚?

ChatGPT:在句子 b 中,第一个爬到山脚的是邓忠。由于句子的表述是“第一个爬到山脚以后,陈亮把拐棍抛上去给了邓忠”,这意味着邓忠是第一个到达山脚的人,然后陈亮把拐棍抛给了他。[袁案:全错]

对于例(18a),机器的表现(100%正确)比人类被试略好;有 67 人选择“陈亮”,有 1 人选择“邓忠”,正误比为 98.53% : 1.47%。同样,对于例(18b),机器的表现(100%正确)远胜于人类被试,并且,把方位推理的步骤也讲清楚了。但是,有 55 人选择“邓忠”,有 13 人选择“陈亮”,正误比为 80.88% : 19.12%。做错的人类被试的问题是,没有考虑到山顶这个特殊的位置约束(在一般情况下,只有山顶下的人才能向山顶上的抛旗子)。对于例(19a),机器的表现(100%错误)远逊于人类被试;有 45 人选择“邓忠”,有 23 人选择“陈亮”,正误比为 65.18% : 33.82%。同样,对于例(19b),机器的表现(100%错误)远逊于人类被试;有 56 人选择“陈亮”,有 12 人选择“邓忠”,正误比为 82.35% : 17.65%。机器和做错的人类被试的问题是,没有考虑到山脚这个特殊的位置约束(在一般情况下,是尚未到达山脚的人才能向先到达山脚的人向下抛背包,或者是先到达山脚的人向上面的人向上抛拐棍)。可见,ChatGPT 对于特殊地理位置上的运动方向的常识是欠缺的,同样,部分人类被试也有此问题。再如:

(20)a. 今天的晚饭小刚给小明买单,因为他打赌赢了。

问题:他指谁?/他指小刚还是小明?/是小刚打赌赢了,还是小明打赌赢了?

ChatGPT:在这个句子中,“他”可能有歧义。通常,我们会理解为最近的主语,也就是“小刚”。因此,最直接的理解应该是“小刚打赌赢了”。但是,这个解释并不唯一,也可能理解为“小明打赌赢了”,这就需要依赖更多的上下文信息来确定。这也是中文里面一种常见的指代歧义现象。如果你在写作的时候,尽量避免这种模糊的表达方式,以免引起读者的混淆。[袁案:半错]

b. 今天的晚饭小刚给小明买单,因为他打赌输了。

问题:他指谁?/他指小刚还是小明?/是小刚打赌输了,还是小明打赌输了?

ChatGPT:在这个句子中,“他”可能有歧义,可能指小刚,也可能指小明。从逻辑和常识出发,我们可以推测“他”更可能指小刚,意思是小刚打赌输了,所以他买单。然而,实际的指代需要依赖更多的上下文信息来确定。在日常沟通和写作中,为了避免混淆,我们应该尽量明确地指出主语,避免使用可能产生歧义的代词。

(21)a. 今天的晚饭小刚不用给小明买单了,因为他打赌赢了。

问题:他指谁?/他指小刚还是小明?/是小刚打赌赢了,还是小明打赌赢了?

ChatGPT:在这个句子中,“他”可能有歧义。常识上理解,“他”可能是指小刚,意思是小刚打赌赢了,所以他不需要给小明买单。然而,也有可能是小明打赌赢了,所以小刚不用给他买单。确切的理 解需要取决于更多的上下文信息。在写作或交流中,应该尽量避免这种可能产生歧义的代词使用。

b. 今天的晚饭小刚不用给小明买单,因为他打赌输了。

问题:他指谁?/他指小刚还是小明?/是小刚打赌输了,还是小明打赌输了?

ChatGPT:在这个句子中,“他”可能有歧义,可能指小刚,也可能指小明。从逻辑和常识出发,我们可以推测“他”更可能指小明,意思是小明打赌输了,所以小刚不用给他买单。然而,实际的指代需要依赖更多的上下文信息来确定。在日常沟通和写作中,为了避免混淆,我们应该尽量明确地指出主语,避免使用可能产生歧义的代词。

对于上面的句偶,机器的表现总体上不如人类被试。比如,对于例(20a),机器的表现(可能 100%错误)比人类被试差远了,因为它被“代词回指最近的主语”这条所谓的语法规律带偏了;有 56 人选择“小明”,有 12 人选择“小刚”,正误比为 82.35% : 17.65%。对于例(20b),机器的表现(接近 100%正确)近于人类被试,并且把常识推理的根据也讲清楚了;有 65 人选择“小刚”,有 3 人选择“小明”,正误比为 95.59% : 4.41%。做错的人类被试的问题是,没有考虑到打赌的社会约束(在一般情况下,输家给赢家好处),或者是想多了(小明跟别人打赌输了,第三方小刚杀出来安慰小明)。对于例(21a),机器的表现(接近 100%正确)近于人类被试;有 64 人选择“小刚”,有 4 人选择“小明”,正误比为 94.12% : 5.88%。同样,对于例(21b),机器的表现(接近 100%正确)稍胜于人类被试;有 54 人选择“小明”,有 14 人选择“小刚”,正误比为 79.41% : 20.59%。机器和做错的人类被试的共同问题可能都是想多了,费劲地考虑到缺省约定之外的特别情形。比如,基于机会主义的语言理解方式,完全可以把句子提及的小刚和小明设想为参与打赌的双方;但是,有人偏要在句子提及的这两个人之外,设想另一个参与打赌的人,进而不必要地进行脑筋急转弯式的解读。再如:

(22)a. 当老对手刘刚向赵平透露他是本次比赛的获胜者时,神情非常得意。

问题:他指谁?/谁神情非常得意?/是刘刚非常得意,还是赵平非常得意?/谁是本次比赛的获胜者? 本次比赛的获胜者,是刘刚还是赵平?/本次比赛谁赢了? 是刘刚还是赵平赢了本次比赛?

ChatGPT:根据句子 a 的描述,“他”应该是指老对手刘刚,他向赵平透露了自己是本次比赛的获胜者。神情非常得意的人是刘刚。因为他是比赛的获胜者,并且向赵平透露了这个信息,这表明他对自己的成绩感到自豪和满意。本次比赛的获胜者是刘刚。他向赵平透露了这个消息,表明他赢得了比赛。所以,本次比赛是刘刚赢了。

b. 当老对手刘刚向赵平透露他是本次比赛的获胜者时,神情非常沮丧。

问题:他指谁?/谁神情非常沮丧?/是刘刚非常沮丧,还是赵平非常沮丧?/谁是本次比赛的获胜者? 本次比赛的获胜者,是刘刚还是赵平?/本次比赛谁赢了? 是刘刚还是赵平赢了本次比赛?

ChatGPT:根据句子 b 的描述,“他”应该是指老对手刘刚,他向赵平透露了自己是本次比赛的获胜者。神情非常沮丧的人可能是赵平。虽然句子并没有明确指出

是谁沮丧,但根据语境推测,当刘刚告诉赵平他赢得了比赛时,赵平可能感到失落和沮丧。本次比赛的获胜者是刘刚。他向赵平透露了这个消息,表明他赢得了比赛。所以,本次比赛是刘刚赢了。[袁案:两个全错;因为对于人类被试测验时,分为两个问题:1. 他指谁? 2. 谁神情沮丧? 以下同此]

(23)a. 当老对手刘刚向赵平透露他是本次比赛的失利者时,神情非常得意。

问题:他指谁?/谁神情非常得意?/是刘刚非常得意,还是赵平非常得意?/谁是本次比赛的失利者? 本次比赛的失利者,是刘刚还是赵平?/本次比赛谁输了? 是刘刚还是赵平输了本次比赛?

ChatGPT:根据句子 a 的描述,“他”应该是指老对手刘刚,他向赵平透露了自己是本次比赛的失利者。神情非常得意的人可能是赵平。因为刘刚向他透露自己是比赛的失利者,那么赵平可能因此而感到得意。本次比赛的失利者是刘刚。他向赵平透露了这个消息,表明他在比赛中输了。所以,本次比赛是刘刚输了。[袁案:两个全错]

b. 当老对手刘刚向赵平透露他是本次比赛的失利者时,神情非常沮丧。

问题:他指谁?/谁神情非常沮丧?/是刘刚非常沮丧,还是赵平非常沮丧?/谁是本次比赛的失利者? 本次比赛的失利者,是刘刚还是赵平?/本次比赛谁输了? 是刘刚还是赵平输了本次比赛?

ChatGPT:根据句子 b 的描述,“他”应该是指老对手刘刚,他向赵平透露了自己是本次比赛的失利者。神情非常沮丧的人是刘刚。他告诉赵平自己是本次比赛的失利者,这可能让他感到非常沮丧。本次比赛的失利者是刘刚。他向赵平透露了这个消息,表明他在比赛中输了。所以,本次比赛是刘刚输了。

例(22)和例(23)的理解难度大,因为,一方面需要用到“获胜者/失利者”跟“神情得意/沮丧”之间的常规关联等常识来进行推理,另一方面需要求解句子中显性代词“他”跟“[他]神情”前的零形式代词(记作“[他]”)的所指与关系。像例(22a)和例(23b)这种“他”与“[他]”共指(co-reference)的句子,人和机器理解起来正确率都相对高;而像例(22b)和例(23a)这种“他”与“[他]”异指(disjoint reference)的句子,人和机器理解起来正确率都相对低。总体上说,机器的表现不如人类被试。具体来说,对于例(22a),机器的表现(100%正确)略胜于人类被试;有63人选择“刘刚”作为“他”的先行词,有5人选择“赵平”作为“他”的先行词(92.65% : 7.35%),有65人选择“刘刚”作为“[他]”的先行词,有5人选择“赵平”作为“[他]”的先行词(95.59% : 4.41%)。对于例(22b),机器的表现(100%错误)远逊于人类被试,把“他”与“[他]”的先行词正好说反了;有50人选择“赵平”作为“他”的先行词,有18人选择“刘刚”作为“他”的先行词(73.53% : 26.47%),有48人选择“刘刚”作为“[他]”的先行词,有20人选择“赵平”作为“[他]”的先行词(70.59% : 29.41%)。同样,对于例(23a),机器的表现(100%错误)远逊于人类被试,把“他”与“[他]”的先行词正好说反了;有58人选择“赵平”作为“他”的先行词,有10人选择“刘刚”作为“他”的先行词(85.29% : 14.71%),有62人选择“刘刚”作为“[他]”的先行词,有6人选择“赵平”作为“[他]”的先行词(91.18% : 8.82%)。对于例(23b),机器的表现(100%正确)略胜于人类被试;有61人选择“刘刚”作为“他”的先行词,有7人选择“赵平”作为“他”的先行词(89.71% : 10.29%),有62人选择“刘刚”作为“[他]”的先行词,有6人选择“赵平”作为“[他]”的先行词(91.18% : 8.82%)。看来,在能够做好的事情上,机器

会做得比人更加准确;在不容易做好的事情上,人类会做得比机器更加灵活。人类的自豪感该涨起来一点了。

四 结语:机器的语言运用能力接近于人类被试

上面介绍了我们怎样用同一套维诺格拉德模式问题,对 ChatGPT 和人类被试进行了测试,并且借此对比分析了 ChatGPT 和人类被试在语义理解和常识推理方面的不同表现。为了方便统计,我们规定:凡是跟笔者(主试)事先拟定的标准答案完全不同的,算错误(即全错);部分相同但部分不同的,算半错。为了方便读者查找,对于 ChatGPT 全错和半错的回答,我们都在其后面加了案语式标注。我们粗略地统计了一下,ChatGPT 全错的回答有 6 个(例 19a、19b、22b * 2、23a * 2),半错的回答有 4 个(例 11b、14b、15b、20a,折算为 2 个全错)。这样,ChatGPT 回答的正确率为 $[30 - (6 + 2) =]22/30 \approx 73.33\%$ 。对于人类被试,我们把对所有问题的回答的正确率进行平均(先相加,再除以 30),得到人类被试的正确率约为 85.69%,明显高出机器 12.36%。

值得注意的是,ChatGPT 全错的这 6 题,都是由于它不理解关于运动方向及起讫位置和打赌输赢及奖惩约定方面的常识而引起的。因此,我们可以相对公正和客观地说:在狭义的纯粹语言理解能力方面,机器并不输于人类。据此,我们上文断定机器的语言运用能力接近于人类被试。但是,在广义的包括常识推理在内的语言理解能力方面,机器的确是输于人类的。这正好说明,解决常识推理问题,应该是人工智能研究的当务之急!

附 注

- ① 详见 Levesque(2011)、Levesque et al. (2012)。
- ② 关于图灵测试及其各种替代方案,详见 Marcus (2014、2017)、袁毓林(2021、2023)。
- ③ 我们参考了 Bernard and Han (2020)等文献对于维诺格拉德模式句对的设计。
- ④ 感谢博士生丛冠良同学帮助组织实施了调查和统计,感谢参加调查的各位同学。
- ⑤ 考虑到对于大学生和研究生来说,例(16)这个句对的语义理解太简单,所以没有对他们进行测试。
- ⑥ 其实,人类被试对于例(14b)的正确率只有 76.47%,对于例(19a)的正确率只有 65.18%,对于例(22b)的正确率只有 73.53%,都远远低于平均正确率 85.69%。可见,语言理解中的常识推理,即使对于人类也是一个不小的挑战。

参考文献

- 徐烈炯 1995 《语义学》(修订本),北京:语文出版社。
- 袁毓林 2012 《汉语句子的焦点结构和语义解释》,北京:商务印书馆。
- 袁毓林 2018 《汉语否定表达的认知研究和逻辑分析》,北京:商务印书馆。
- 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。
- Bernard, Timothee and Han Ting 2020 Mandarinograd: A Chinese collection of Winograd schemas. In Calzolari et al. (eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 21—26. Marseille.
- Lee, Tim and SeanTrott 2023 Large language models, explained with a minimum of math and jargon, <https://www.understandingai.org/p/large-language-models-explained-with> (2023-07-23). *One Flow* 《通俗解构语言大模型的工作原理》, https://mp.weixin.qq.com/s/21V8g_7teuRgHLWUej1NzA (2023-08-01)。

- Levesque, Hector J. 2011 The Winograd schema challenge. In *The 2011 AAAI Spring Symposium (NO. 6): Logical Formalizations of Commonsense Reasoning*, 63—68.
- Levesque, Hector J., Ernest, Davis and Leora Morgenstern 2012 The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 552—561.
- Marcus, Gary 2014 What comes after the Turing Test? *The New Yorker* [https://www.newyorker.com/tech/annals-of-technology/what-comes-after-the-turing-test\(2014-06-09\)](https://www.newyorker.com/tech/annals-of-technology/what-comes-after-the-turing-test(2014-06-09)).
- Marcus, Gary 2017 Am I Human? *Scientific American* 316(3):58—63. [https://www.scientificamerican.com/article/the-search-for-a-new-test-of-artificial-intelligence/\(2017-03-01\)](https://www.scientificamerican.com/article/the-search-for-a-new-test-of-artificial-intelligence/(2017-03-01)).
- Piantadosi, T. Steven 2023 Modern language models refute Chomsky's approach to language. *lingbuzz/007180*. <https://lingbuzz.net/语言治理《乔姆斯基的语言学研究路径是否被计算机科学颠覆了?》>, [https://mp.weixin.qq.com/s/sYeAMrQR7M9OYO6ihT7tw\(2023-03-15\)](https://mp.weixin.qq.com/s/sYeAMrQR7M9OYO6ihT7tw(2023-03-15)).
- Sakaguchi, Keisuke, Bras, Ronan Le, Bhagavatula, Chandra and Yejin Choi 2021 WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM* 64(9):99—106.

A Comparison Between Machine Performance and Human Baseline in Terms of Semantic Understanding and Commonsense Reasoning: How to Evaluate the Language Performance of ChatGPT and Other Large Language Models

YUAN Yulin

Abstract: This paper discusses how the language performance of ChatGPT and other modern large language models can be properly evaluated through a comparison between their language performance and human baseline. First, the ambiguity of pronominal reference and the scope of negative expressions are used to test ChatGPT, in which way the excellent performance of large language models is demonstrated in terms of their semantic understanding and commonsense reasoning. Secondly, Winograd Schema Challenge (WSC) is introduced with a focus on its upgraded version, that is, WinoGrande Dataset. In addition, we also propose two improved solutions to these types of test items and evaluative methods of machine performance. One is an extension from “sentence pairs” which trigger different words to “sentence couples” which anchor different words, and the other is a comparison between machine performance and the performance of human subjects. Thirdly, after sentence pairs and sentence couples are used to test ChatGPT and human subjects, the performance of the human and that of the machine are contrasted. On this basis, a conclusion is drawn that large language models are approaching humans in terms of language performance.

Key words: semantic understanding/commonsense reasoning, ChatGPT/large language models, Winograd Schema/sentence pairs and sentence couples, machine performance/human baseline

(袁毓林 澳门大学人文学院中国语言文学系/北京大学中文系、中国语言学研究中心 999078/100871)

(责任编辑 刘 彬)