第三届“卿云杯”全国通识课程论文大赛

姓名	朱子萱	**院系专业**	经济管理试验班
学号	21307100110	**任课教师**	陶晓鹏、危辉
课程名称	从计算到智能		
论文题目	从华语辩论与机器辩论的鸿沟看不成熟的机器辩手		

从华语辩论与机器辩论的鸿沟看不成熟的机器辩手

摘要：机器辩论是人工智能的一个新兴领域，它综合利用了人工智能领域的许多工具与方法，获得了一定行业认可。然而，从技术科学角度看，其算法仅实现了表面流程化的辩论话术表演，并未从辩论的本质上实现逻辑的推演。从辩论内涵角度看，机器的论辩缺乏灵活性与流畅性，亦无法进行哲学命题的思辨。因此，机器辩论被高估了，机器辩论没有还原真正意义上的辩论。
关键词：机器辩论；人工智能；自然语言处理

机器可以证明数学定理，那机器能否论证社会议题？机器可以赢下围棋冠军，那机器可否击败辩论之王？为辩论而生，IBM 的 AI 系统 Project Debater 是第一个机器辩手。它出道即巅峰，首战便击败国家辩论冠军，可堪辩论界的 AlphaGo。辩论，是非常人文的竞赛活动。机器在这一场域的崛起，是否宣告着 AI 终于得以跻身谈经对饮的高雅殿堂？我以为不然。在观摩了这些人机唇枪舌战后，我愿断言：机器辩论距成熟仍任重道远。并且，其目前的几场胜局代表不了什么，华语辩论赛场将会是机器辩手的滑铁卢。

本文从一位业余辩论爱好者的视角出发，试图分析机器辩论为何依旧低阶，并尝试指出目前形态下的机器辩手和人类辩手之间那道难以逾越的鸿沟。笔者希望借助本文为大家看待机器辩论提供一个更冷静理性的视角。

（注：Project Debater 诞生于 2018 年。直到今天，它都是世界上唯一的机器辩手。我们只好姑且以它代表目前机器辩论的水平，分析它的行为范式来得出今天的结论。）

一、复盘：赢下比赛的 AI 辩手

看完这些比赛，笔者愿意把票投给 AI 辩手，但这并不代表笔者承认这是 AI 里程碑式的进化。为何言此？理由如下。

首先，观众们必须认识到，在这些比赛中，人类的辩论综合素质（攻防意识、临场反应）都明显优于机器辩手。但是，机器的确做到了利用更充足、有效的论据进行反驳，而这些直接奠定了它的胜局。在很多回合，笔者都能看出人类辩手有明显的反驳意识，只可惜其论据较为单薄，以至于只能不断重申应然层面的价值方针，无法拿出驳倒对方的见血论断。而机器却得以游刃有余地使用权威数据、专家方案来一步步推逻辑，最终完成逻辑闭环，建立牢不可破的防线。再精锐的特战员，若只有一把锄头，在拿火枪的流匪面前也甘拜下风。这一役，人类纯粹是吃了装备不足的亏，而机器也不过只是乘了论据的势。

其次，我认为这些赛程的设置有失偏颇。在这几场比赛中，辩题都是当场下发给双方辩友的。虽然主持人一再强调，机器面临的是它从未训练过的辩题、机器只有很短的准备时间云云，但是，计算机的数据处理速度是凌驾于人类之上的。

索引、关联评估、拼接文本……机器可以多项任务并行处理；人类辩手则只得全凭一己之力，逐条查阅，手工记录，且比赛开场后便再无机会针对对方抛出的论点靠外界资源搜索反面论据了。让双方分别面对开卷考试和闭卷考试，这场对决的公允性很难不被质疑。但凡让人类备赛的时间稍长一点，机器辩手便多半南风不竞了。

此外，Project Debater 与世界冠军 Natarajan 进行那场比赛的严谨性也是值得质疑的：Natarajan 似乎在比赛中放了水。在一个回合中，Project Debater 误解了 Natarajan 的论点，随后试图反驳这个对方甚至没有提出的论点。Natarajan 本可以揪住这个漏洞穷追猛打，不过，这位人类展现了君子之风，只是继续阐述他的论点，装作没有注意到对方的错误。这可能是因为 Natarajan 想尽可能多地和机器进行更有效的交流，而不愿耽误时间在细节上纠缠。但这也正说明了，这场比赛并没有贯彻国际赛事的水准，更像一场请冠军来助个兴的发布会或表演赛。它与 AlphaGo-李世石那场双方毫无保留实力的生死对决的意义是不可相提并论的。

最后，令人哭笑不得的是，主办方在辩论结束之际发起了一项快速投票调查。几乎毫无疑问地，调查的结果显示，69%的观众都认为“Project Debater 比人类的论述更能丰富他们的知识”。不得不说，如果主办方诚心想实事求是地知道自己的机器辩手到底是否出色，就该换掉这种倾向性过于明显的问题。我相信，我们在乎的不是知识的输出，而是能力的展现。

由此观之，这样赢下比赛，真的像 IBM 所宣称得那样有很大意义吗？在我看来，这分明是高级版的应用文创作器摇身一变，就贴上了“辩论”的华丽标签。我们已见识了机器辩论在赛场谈不上精彩的表现，接下来，让我们从舞台走向幕后，看看 AI 辩手是如何养成的吧。

二、尴尬的训练战术

为了呈现出有血有肉的陈述，AI 辩手的首要任务是建立自己的素材库。面对浩如烟海的资料，机器整理分类的方式直接影响了其后期填充讲稿的架构。大约是为了给讲稿写作时的调用做铺垫，研究人员让 Project Debater 在数以亿计的论文、报告和新闻文章中按（1）观点（2）证据 的分划来整理素材。在得到素材后，机器就要套用模板了：将分论点和支持它们的论据放入事先准备好的文章框架中。

研究人员重点针对这一过程对机器进行反复训练，使这些素材的组合尽可能强大而贴切。这意味着，要去确保这些论点和证据存在紧密的逻辑顺承，以及确保这些论点足够见血而有洞见、这些证据权威且相关等等。不过，AI 的训练内容也仅仅着重于此，停留于修复优化这些表层细节了。机器辩论的训练并未涉及

教会机器自己利用数据总结结论。笔者大胆揣测一下，这或许是因为，通常外行人看到“机器辩论”，只会关心它说了什么，而不关心这些话语是从哪里来的——是他人之言或是自己所想。这类肤浅的看法让机器和研究人员倾向于选择偷懒：既海量文献已给出解答，不妨直接伸手摘取那现成的果实？然而，这二者内在有质的区别。若是他人之言，那讲得再精巧有洞见，也不过是从拼凑缝合升级到了流畅转述；若是自己所想，那纵使这位机器辩手磕磕绊绊、观点稚嫩，我们也可以说，我们看到了人工智能向前的一步。可惜，AI 辩手目前的训练内容都属于帮助其发言“从拼凑缝合改进为流畅转述”这一范畴，而对后一领域的探索并未得到重视。

“授人以鱼，不如授人以渔”，重复训练机器更精准地抓取有用信息，不如试着锻炼其自己用数据归纳形成结论。毕竟，前者已经有很多人在做了，也已经有很多人做过了；而后者才是我们最需要突破的地方，亦是机器辩论所能做到给“辩论”二字的最佳诠释。

然其训练之问题不止于此。比上述更尴尬的是，Project Debater 在训练时接触到的材料，仍然需要细致的人工标记。这几乎是人类在手把手教学：如果机器不能突破自己识别检测观点、事实，给话题分类，鉴别强论据和弱论据，那么机器便无法独立准备一场辩论赛。

当然，这也有情可原。对机器来说，辩论之所以比围棋、纸牌游戏更难，就是因为辩论的结局是开放的，没有像“围出更多空点”或“打掉手牌”这样具体的胜利条件。机器在陈述完一段观点后，由于无法获得反馈，它永远不会自己意识到方才所用的说服是否是强悍的、反驳是否是绵软的，如此等等。

因此，研究人员不得不用大量手工打好标签的高质量数据，投喂、饲养 AI 辩手，以便用弱监督的方式，以深度神经网络（Deep Neural Networks）训练它。

如果能让人类做到适当放手，令机器做到自己去分辨概念，自己去识别那些宽泛的领域，自己去推理，机器就能从原理上（而不是表观上），更靠近人类辩手。可惜人类目前还无法撒手。

三、论辩的真谛和机器的拙劣模仿：比较的缘由

至此，想必对机器辩论智能乏新的揭露，已不必再赘述。是时候驶离算法，谈谈真正的辩论了。为什么我一定坚持要比较机器辩论在 IBM 主导的赛制和华语辩论中可能的不同表现，又近乎严苛地要求机器去挑战华语辩论的擂台？我想指出的是：我们之所以认为机器能辩论是一件了不起的进步，就是因为辩论这一活动天然地具有着智识、思辨、沟通这些人类的特征。当聊起辩论，我们脑中浮现的是古希腊集会上激情演说的雄辩家，是春秋时期游走于刀光剑影的纵横家，是老电影中法庭上主持正义的刑辩律师，是脱口秀里引爆全场的谐星演员。IBM 的

发明引导我们把以上形象与人工智能联系在一起，进而惊叹人工智能发展成绩斐然。但如果在实际辩论中机器并没有很好地展现出这些特征，甚至相反地去回避这些高难度谋术，那么，在击败冠军的光环下，这个 AI 真正的突破点在何处，是值得我们理性反思的。

请让我解释 IBM 设置的辩论和华语辩论的巨大区别。它们分别代表两种辩论类别——Project Debater“擅长”的政策辩和华语辩论中常见的价值辩。政策辩起源于英国议会辩论，因此又称议会辩论，顾名思义，其辩题多与社会政策相关；而价值辩，则是叩问人心中价值排序的哲理之争。

下面是 IBM 为训练机器辩手所建辩论数据库中的辩题。从这些高度趋同的措辞可以看出，这位西方机器打的全是政策辩。

IBM 辩论数据库中辩题

The use of trans fats in food **should be banned**
Software patents **should not be allowed**
Smoking **should be further restricted**
Clean energy **should be employed**
The monarchy **should be abolished**
The sale of violent video games to minors **should be banned**

相对地，华语辩论的辩题常遵循先验的原则，关注应然和价值。以下为一些华语辩论圈中公认知名赛事的辩论题目，其风格与政策辩之迥异可见一斑。

华语辩论知名赛事辩题

乱世中，人们应有宁为玉碎，不为瓦全的精神/乱世中，人们应有留得青山在，不怕没柴烧的思想（2004 亚太大专华语辩论赛）
自由意志存在/不存在（2016 年新国辩决赛）
被同化/被排斥更可怕（2017 华语辩论邀请赛）
人生若只如初见，是可喜的/可悲的（2021 华语辩论世界杯）

这四个辩题分别体现了华语辩论辩题设置中极端或虚拟场景的预设、高度抽象化的表达、哲学概念强烈的矛盾感、基于个体生命境遇的思考。荆天棘地如此，光是破个题可能就会让我们的机器选手 CPU 发烫了。面对这些辩题，机器能否故伎重演，诉诸其强大智囊——万能的数据库来获得力量呢？下面我将根据 Project Debater 的各项表现，揭示它在华语辩论面前的无力。

四、推逻辑之外的辩论

以下四个特征乃华语辩论的代表性特色，在议会辩论中则少有涉及。正是它们，让华语辩论有了观赏性和魅力，把“辩论”又向靠近人的地方推了一步。

Project Debater 没有表现出这些特征，这决定了它充其量为一归纳资料的

顾问助理，而不是真正学究天人的辩才。

（一）没有退路的对攻与交锋

IBM 组织的辩论赛共尝试了两种赛制，第一种赛制让人类与机器分别进行四分钟开场演讲，四分钟反驳和两分钟论证总结；第二种赛制要求机器先陈述三分钟，人类陈述三分钟，共计三回合。

较长的陈词时间对机器极度有利。首先，这样机器加载发言的时间更宽裕。语音识别、文字转换、关键要素提取和检索整理信息都意味着工作量，IBM 也承认，机器需要不小的时间去完成这些步骤。其次，面对人类长达 700 单词左右的辩词，机器并不需要逐点反驳，只需抓住主要矛盾或有相关反例支撑的点去驳论即可，而这纵容了一些计算机的模糊化处理。

相比之下，华语辩论的赛制则一般涉及自由辩论和 1v1 攻辩环节。在这些环节中，双方需要在十几秒内正面回应对方刚刚抛出的问题，并反诘对方，快速推动赛程进展。如此一来，机器的准备时间大为缩短，且对其回应内容的质量要求显著增加。三五句话的限制下，要求机器做到反驳精准、到位，拆论见血、有力，这是自然语言处理前所未有的挑战。

在对战 Natarajan 那场的“反驳”环节，机器甚至还能颇有“闲情逸致”地朗读大段含有具体数字的数据、介绍一些现行政策。在观众看来，这可能是“例证充分”、“证据有力”，然而，这个人工智能恰恰只是做不到提炼概括出精炼的表达，只能被迫忠实地复述别人说过的话。

（二）辩才无碍：无巧舌，不辩论

说服的艺术不只是推演逻辑。若是那样，双方根本没必要发出声音，只需把尽可能多而强的论据和分论点逐条列出，提交至评委处即可。

华语辩论尤其强调语言的优美性、句式的节奏感和语势的起伏，这些都是辩手在呈现观点时需要考虑的。以排比句为例，我们可以看到，尽管 IBM 设置的辩论赛本身文学性不强，Natarajan 也至少尝试使用了若干排比句（it is terrible that……, may be it does……）来提升表达质量。而对于 Project Debater，它全场下来使用的排比句数量是零。

我们能看出 Project Debater 在提高流畅性上做出的努力，比如试图往材料之间塞关联词（next, furthermore, in addition, finally, etc.），这涉及了句与句之间的处理，或者段落与段落之间的处理。然而，对于一个句子内部的加工，我们几乎看不到。这暴露出，机器难以进行句子内部（即单词与单词之间）手术刀式的精细化操作。首先，这应归咎于尚不成熟的自然语言处理技术，在自然语言精修技术的欠缺下，任何改动原句句式的企图都可能导致语法错误，甚至致命的表意扭曲。

进一步地，这也折射出了机器辩手沟通意识的缺乏。它忙于闷头梳理论据、拼装论点，而忽视了“说服”这一动作是有具体对象的。为了打动听众，一定要有对话的意识、交流的意识，而这些均体现在非常细微的用词上，比如恰到好处的反问、引起听众兴趣的互动。所有这些动作，人类都可以在潜意识下轻松连贯地完成，而机器则必须非常刻意，才有可能勉强显露出少许这样的语言氛围。

拥有知识、概念和观点是容易的，自然语言生成、表达型文本转语音也造成不了太大的挑战，这里的难点在于如何驰骤殿庭、周旋击拂、曲尽其妙地完成素材的表达。这是口语的艺术、交流的艺术，而不是写作的艺术。令我啼笑皆非的是，Project Debater 的陈词于我而言，有一种强烈的熟悉感——它像极了我写出来的托福考试作文：标准得不能再标准的三段论的模板、观点例子层层相叠的拼装组合，如此再在开头加个开场白，结尾补上总结句，中间不忘过渡语，一篇陈词就完成了。的确，它结构严谨，思维缜密，但我看不出这其中口语化的表达体现在哪里，我看不出交际性在哪里，我看不出它说服听众的强烈意图……这不是辩论，而我们不想仅仅要一个标准化稿件生成器。

（三）机器克星：类比推理的不按套路出牌

如果说，以上两点都是可以通过优化算法改善的，那么接下来的问题便不是可以那么轻易解决的了。至此，我们已触及到了人和机器间那条边线。

让辩词有说服力的一个方式是使用精巧的类比。这和统计数据、证据无关，是一种概念和思维上的推演。Project Debater 并没有使用这种高级思维。

试举一简单例子，要证明“知易行难（想到的不一定做得到）”，千言万语都不如“谁都知道圆是什么形状，但是谁都不能用手画出一个正圆”有冲击力。类比，可谓四两拨千斤。用简单形象的类比说明一个复杂的问题，自古都是人类沟通中常见而有力的说服手段。

当人类的大脑试图理解相似之处时，我们更关注“关系”而不是“特定对象”。而这给了人类一个伟大的思维模式：类比思维。无论是物理中水流模型和电流模型的精妙类比，还是压根不存在什么逻辑的“人性之善也，犹水之就下也”式的臆想类比，都是人类在动用一种非常智能的思考方式。

而这种人类观众易于接受的自然流畅的类比推理，在机器看来却可能完全不直观。在三亿文献中搜寻有用论据时，机器辩手只能发觉“对象”的关联，而无法发现“关系”上的关联。在选取论据时，机器搜寻信息的原理和论文查重大同小异——抓取重叠信息。这样的扫描、排查利于快速对应到针对性的信息，如处理“Smoking should be further restricted”这个短句时，机器只要在数据库中搜索“smoking”和“restrict”，再配合加权一些与“支持”“认为”相关的语汇，则其最终搜寻到的观点、论据都不会离题太远。句级索引和语义关系评估

是机器的看门本领，但是，这也有一个致命的缺陷：机器只能追查到关键词及其近义表达，而对于那些支持本方论点的潜在证词，机器很难拿来为己所用。

不解决机器在概括和抽象能力上的欠缺这一问题，机器便永远无法把说理拔高到关系的维度，道出思辨层次的洞察和见解。

（四）我口说我心

最后，华语辩论中很多关于哲理和生活的讨论，是没有生命体验的机器辩手难以回应的。“纸上得来终觉浅，须知此事要躬行。”不巧，机器偏就缺乏了这关键的生命体验一环。硬逼着它发表看法，它也只能试从哲学学术论文中寻章摘句，强作解人了。

且不说机器能否整合思想界纷杂流派的各路观点而不失其要领，光是给机器出一道哲学思辨类的辩题，就无异于让 AI 辩手在“中文房间”中扮演一位久经世事的哲学家。而这样一位扮演出来的哲人只能处理语言、整合数据，它并不确定它所捍卫的论点背后的道德依据是什么（当然，我们目前大抵也不会尝试让 AI 进行道德推理，且不论其推理产物到底能否代表权威，从伦理角度来看，这也是很危险的）。

辩论，是为了启发思考。说得崇高一点，辩论是在要求辩手为思想界“贡献”出自己独特的想法和见解，给听众、给决策者提供更深或更广的洞见视角的。如果辩手自己拿不出任何东西，只是复述他人的成果，那他在场上发言这一行为本身的意义就被消解了。如果人们认为让机器有想法是难的，那至少也让机器做一些简单的判断推理吧。机器辩论若不跨出这关键的一步，就永远只能在辩论场外围徘徊。

五、有限的成就

在展演的末尾，IBM 交代了其关于机器辩论未来的期望。IBM 表示，他们希望这项技术可以支持政府政策拟定、企业产品回馈等方面。的确，这项技术的前景是广阔的——它可以应用于（1）需要研究、分析、收集信息的领域，比如金融分析、新闻；（2）需要做决策的领域，比如公司战略管理。在笔者看来，对一个集强大搜索引擎和一定信息交互能力于一身的 AI 来说，胜任这些不在话下。

问题在于，这些技术我们在很多年前就已经实现了。它的创新点在哪里？论信息处理能力，它不是领衔级的；论言语交流能力，它不过比语音助手稍强，但也绝不敌那红极一时的“公民机器人”索菲亚。

笔者相信，当人们最初被这个新事物吸引的时候，期待的绝对不止是企业服务之流。既然 IBM 的研究团队已经打出了智能辩论的旌旗，就应当把向“论辩”高峰吹响号角的勇气贯彻到底。关于机器辩论，还有太多更有意义的问题等待着研究人员去发掘、解答。Project Debater 这个牙牙学语的孩子，何时释卷，夺

席谈经？AI 辩手和我们都有浩浩的长路要走。

就像在阅卷老师面前，给作文自动打分的人工智能无法越俎代庖；在辩论场上，Project Debater 也不能替代辩才之魂。

距 Project Debater 问世，已经过去了三个春秋。这期间，人工智能仍不断在不同领域活跃着，翻新推出着各类花样的产品，但第二代 AI 辩手却一直没有出现。大约，短期内它是不会现身了。

AI 和苏格拉底的距离，依旧天悬地隔。

造势者的口角春风，唤不起人工智能的寒冬。

也许，我们也不需要第二个 Project Debater。唯变革性的硬科技才是那一阵消融冰雪的信风。