清华大学孙茂松：自然语言处理一瞥，知往鉴今瞻未来

孙茂松 AI科技评论 2022-03-13 12:08

近日，清华大学人工智能研究院常务副院长孙茂松教授亲手执笔，对自然语言处理的贡献、当前境界与未来挑战进行了深入的探讨。AI科技评论编辑组深有同感，认为此文十分值得一读，故分享之。

作者｜孙茂松

来源｜中国人工智能学会

人类语言（即自然语言）的重要性无论怎么讲都不为过。社会生物学之父爱德华·威尔逊曾说过："语言是继真核细胞之后最伟大的进化成就。"科普畅销书《信息简史》的作者詹姆斯·格雷克也深刻地指出："语言本身就是人类有史以来最大的技术发明。"这些断言带有科学哲学的意味，反映了现代人类对语言本质理解的不断深化。

众所周知，语言是人类所独有的，是思维的载体，是人类交流思想、表达情感最自然、最深刻、最方便的工具。其中这几个"最"字非同小可。语言之于人类就如同空气之于生物，它时时刻刻、无声无息地融通于我们生活的世界中；它是如此的自然以至于我们常常意识不到它的存在，但一旦没有了它，人类将举步维艰。很不幸，人类语言能力正是现代计算机系统所不具备的，呈现出整体性缺失。一个显而易见的逻辑是，没有语言能力的机器，不可能有真正的智能。

自然语言具有无穷语义组合性、高度歧义性和持续进化性等，机器要实现完全意义上的自然语言理解，"难于上青天"。自然语言理解（一个退而求其次的提法——自然语言处理），因其兼具无与伦比的科学意义与学术挑战度，吸引了一代代学者殚思竭虑、前赴后继。

1

NLP对世界人工智能发展的三个里程碑式贡献

"却顾所来径、苍苍横翠微。"笔者认为，自然语言处理研究（包括文本处理和语音处理两个相辅相成的方面）在世界人工智能发展史上有三个里程碑式的"开风气之先"贡献。不揣孤陋寡闻，一孔之见，不一定对，抛砖引玉而已。

第一个里程碑式贡献

现代意义的人工智能技术研究发端于自然语言处理。对机器智能的痴迷与摸索由来已久，1946年第一台通用计算机ENIAC面世，无疑是一个历史分水岭。早在1947年，时任美国洛克菲勒基金会自然科学部主任的 Warren Weaver，在写给控制论之父维纳的一封信中就讨论了利用数字计算机翻译人类语言的可能性，1949年他发布了著名的《翻译》备忘录，正式提出机器翻译任务并设计了科学合理的发展路径（其内容实际上涵盖了理性主义和经验主义两大研究范式）。1951 年以色列哲学家、语言学家及数学家Yehoshua Bar-Hillel在麻省理工学院便开始了机器翻译研究。1954年Georgetown大学与IBM合作的机器翻译实验系统进行了公开演示。机器翻译是典型的认知任务，显然属于人工智能领域。

第二个里程碑式贡献

自然语言处理在人工智能领域乃至整个计算机科学与技术领域较早提出并系统性践行了非结构化"大数据"理念，整体上实现了理性主义研究范式向经验主义研究范式的嬗变。下面举两个典型工作。

一是连续语音识别。自上个世纪70年代中期开始，著名学者Frederick Jelinek领导的IBM研发小组即提出了基于语料库n-gram语言模型（实际上就是n阶马尔科夫模型）的大词表连续语音识别方法，使语音识别的性能上了一个大台阶。这个思路对语音识别领域产生了20年左右的深远影响，甚至包括90年代推出的开创了机器翻译新格局的IBM统计机器翻译模型（该模型使机器翻译研究回归到1949年Warren Weaver建议的经验主义研究范式下，充分展示了他的先见之明）。

二是词性自动标注。1971年曾有学者精心设计过一个TAGGIT英语词性标注系统，使用了3300条人工编制的上下文敏感规则，在100万词次的布朗语料库上获得了77%的标注正确率。1983-1987年间，英国兰开斯特大学的一个研究小组另辟蹊径，提出了不需要人工规则的数据驱动新方法，利用已带有词性标记的布朗语料库，构造了基于隐马尔科夫模型的CLAWS英语词性标注系统，并对100万词次的LOB语料库进行词性自动标注，正确率一举跃升到96%。

第三个里程碑式贡献

当前这一波席卷全球的人工智能高潮肇始于自然语言处理。2009-2010年间著名学者Geoffrey Hinton与微软邓力博士合作，率先提出了基于深层神经网络的语音识别方法，使得语音识别的性能突破了近10年的瓶颈制约，更上一层楼，令学界初步体会到了深度学习的威力，信心顿增，一扫对深度学习框架半信半疑之状态，其后各研究领域遂从者如云，争先恐后如过江之鲫。2016年谷歌推出了深层神经网络机器翻译系统GNMT，彻底终结了IBM统计机器翻译模型，翻开了新篇章。

— 2 —

基于深度学习的NLP：目前形成的基本态势

自2010年以来，深度学习异军突起，日新月异，强力推动了人工智能的全面发展。10年发展的结果是：一方面，深度学习使人工智能技术从几乎完全"不可用"走向了"可用"，取得了历史性的非凡进步；另一方面，虽然它使得人工智能系统在几乎所有经典任务上的性能表现均得以明显提升，但受囿于深度学习方法所存在的深刻短板，在很多应用场景尚达不到"能用、管用、好用"。自然语言处理领域基本上也是这样，本文不赘述。

宏观上看，人工智能领域的发展无例外地得益于两大类型的方法利器：针对图像的卷积神经网络（CNN），以及针对自然语言文本的循环神经网络（RNN）。最初两三年前者风头尤劲，近些年后者贡献更为卓著。若干影响深度学习全局的主要思想，如注意力机制、自注意力机制、Transformer架构，均出自后者。

基于深度学习的自然语言处理，在短短10年中即完成了模型框架上的三次华丽迭代，"从山阴道上行，山川自相映发，使人应接不暇"，先后达至三重境界（实际上这也是深度学习的三重境界）。

第一重境界

针对每个不同的自然语言处理任务，独立准备一套人工标注数据集，各自几乎从零开始（常辅以word2vec 词向量），训练一个该任务专属的神经网络模型。其特点我称之为"白手起家 + 各家自扫门前雪"。

第二重境界

首先基于大规模生语料库，自学习、无监督地训练一个大规模预训练语言模型（PLM），然后针对每个不同的自然语言处理任务（此时也称作下游任务），独立准备一套人工标注数据集，以PLM为共同支撑，训练一个该下游任务专属的轻量级全连接前馈神经网络。在这个过程中，PLM的参数会做适应性调整。其特点我称之为"预训练大模型+大小联调"。

第三重境界

首先基于极大规模生语料库，自学习、无监督地训练一个极大规模的PLM；然后针对每个不同的自然语言处理下游任务，以PLM为共同支撑，通过少次学习（few-shot learning）或提示学习（prompt learning）等手段来完成该任务。在这个过程中，PLM的参数不做调整（实际上由于模型规模太过庞大，下游任务也无力调整）。其特点我称之为"预训练巨模型 + 一巨托众小"。

这三重境界， 一重比一重来得深刻；一重比一重有更多的" 形而上" 感 觉。在GLUE和SuperGLUE公开评测集上的性能表现，也是一重比一重要好（目前正处于第三重）。

近年来，在世界范围内人工智能界各路英豪围绕预训练语言模型展开了巅峰对决，模型规模急剧膨胀（如 2020年6月OpenAI推出的GPT-3模型参数规模达1750亿个，2021年10月微软和英伟达联合推出的MT-NLG 模型飙升到了5300亿个参数），你争我夺，你争我赶，好不热闹。2021年8月，斯坦福大学专门举办了两天的学术研讨会，将第三重境界中的"预训练巨模型"命名为"基础模型"（foundation model），并随即发表了一篇数百页的长文，全面阐述其观点。文中绘制了一张示意图（见图1），揭示了"基础模型"在智能信息处理的中枢作用（其作用疆域已扩展至全数据类型和多模态）。

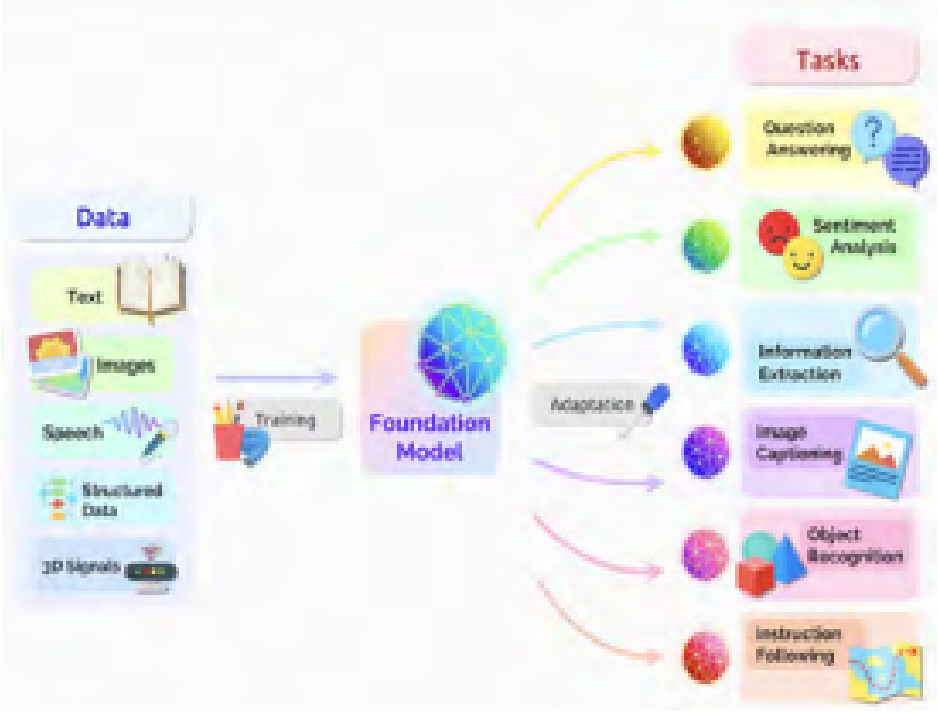


图 1 "基础模型"在智能信息处理的中枢作用

对"基础模型"也有很多质疑的声音，如图灵奖得主Judea Pearl就在推特上发问："'基础模型'使得我们可以绕开以数据为中心的方法的理论局限性的科学原理是什么呢？"(What is the scientific principle by which'Foundation models'can circumvent the theoretical limitations of data-centric methods as we know them...?)

但无论如何，"基础模型"作为智能信息处理公共基础"软"设施的重要地位似已可以基本成立。第一重境界好比各家在海拔几乎为零的沙滩上盖房子，只能自家打好自家的地基，一砖一瓦自己盖起来，很是辛苦，效率也就比较低；第二重境界则是搭建了一个海拔二三十米的共同台面，各家都在这个台面上盖房子，自己打地基的工作大体可以省掉，效率大为提升，但由于台面还不够高，基础不一定处处可靠，所以各家盖房子时还要照顾到台面情况，适当夯实、调整一下台面；第三重境界把公共平台建到了二三百米，地基很牢固，台面高了，即使简单搭个小凉亭，登高望远风景也会不错，各家盖房子需要操心的因素自然就更少了许多，效率得以进一步提升。这是我对"基础模型"定位的一种形象解释。

我们不妨以参数规模达千亿个的百度文心NLP大模型（https://wenxin.baidu.com/wenxin/ernie） 为例，初步体验一下大模型。首先简单考察其"自由创作"功能。随意输入如下一段话，让机器续写：

沿着曲径一路走去，未见

机器续写的句子相当精彩：

到一个人，只有几株高大的银杏树和一些不知名的花朵在阳光下灿烂地开放，偶尔还能够见到两只小鸟。

还是那句输入，让机器再续写一段：

任何人烟，却有一个小山村静谧地伫立在那里。这就是位于大兴区榆垡镇的后石坞乡前石坞村，全村共300户左右、800多口人。如今，农村已经成为许多家庭主要的生活来源之一了。现在我们不仅吃穿住都好，还能看到很多漂亮精彩的节目！

这段话虽然通顺，但文字发展脉络已经有些奇了。

再考察其“自由问答”功能。提问：

山东哪一座山最高？

机器回答：

泰山。

换个问法儿再问机器：

山东最高的山是哪座？

机器回答：

泰山！

考虑到这是生成式回答，机器能做到这一点相当不错。但同样的问题多问机器几次，机器就“蒙圈”了，胡乱回答：

青岛的崂山。这个不用想
五莲山，海拔1095米
济南

可见大模型有比较广博的知识能力，但博而不精，严重缺乏自知之明。所以很容易在语义逻辑上被“攻击”或者“为难”。掉过头来再考察一下其“自由创作”功能。输入：

沿着人来人往的曲径一路走去，未见

机器续写：

一个行人，只有几盏灯笼照着地上的影子。

输入：

> 沿着人满为患的曲径一路走去，未见

机器续写：

> 任何一个人，这里是否有着另外的生命存在？沿途所见皆为枯死之树和倒伏的草木。而他们却不曾注意到身后那片绿意盎然的山谷！“好美啊～”她轻盈跳跃在空中，看向下方正在忙碌的众人道了声赞叹。听闻此言时莫云殇也停止了手上的活计转头望去，只见远处的林间开满了

由上述几个例子可以看出，大模型有明显的长处，但也存在明显的短处。它就是一个有一定海拔高度的台面，在上面可以比较便利地盖房子，但它并不意味着房子已经盖好。其实要在其上盖起一座真正“能用、管用、好用”的房子，还是需要下大气力的。

— 3 —

第三重境界带来的困惑与思考：NLP今后若干年的主要挑战

我认为，自然语言处理在今后若干年主要面临两大挑战。

第一个挑战：力大者为王?

这是顺着目前形成的基本态势，也就是“基础模型”自然向前延伸的。虽然这一路依然在高歌猛进，但“基础模型”本身显然不能解决深层次语义问题；换言之，它注定不能直接导致通用人工智能（虽然它肯定会是设想中的通用人工智能的一个基础部件）。那么，一个自然而然的问题就产生了：这种追求规模（数据、模型、算力）极致化的策略还能走多远？连带的另一个问题是，我们应该做些什么?

我想可以从“基础模型”的开发性（exploitation）和探索性（exploration）两个角度予以回答。

开发性（exploitation）更多注重“基础模型”的工程性，有如下几点应予注意。

- 目前构造及使用“基础模型”的算法本身还是偏粗放型的。前文给出的百度文心 NLP 大模型表现的一些“毛病”，可望通过积极改进算法部分地予以解决。
- 对少次学习、提示学习、基于适配器的学习（adapter-based learning）等与“基础模型”配套的新手段的研发工作应予加强。
- 训练数据包罗万象一定就好吗？是否应对大数据中明显存在着的大量噪声进行筛选?
- 排行榜对模型研发无疑非常重要。但排行榜不是唯一的金标准，应用才是最终的金标准。
- 研发“基础模型”的企业不能“王婆卖瓜，自卖自夸”，要开放给学术界测试。不开放给学术界测试的“基础模型”，其性能是存疑的。学术界不宜盲信盲从。
- “基础模型”亟需找到杀手级应用，才能令人信服地证明自己的能力。

探索性（exploration）则更多注重“基础模型”的科学性。鉴于“基础模型”确实呈现出了一些令人惊奇（或者“奇怪”）的现象，目前尚未给出科学的解释。典型如：

- 为什么大规模预训练语言模型会出现deep double descent现象（这一点似乎超越了机器学习中“数据复杂度与模型复杂度应基本匹配”的金科玉律）?

- 为什么“基础模型”具有少次学习甚至零次学习的能力？这些能力是怎么获得的？其中是否出现了复杂巨系统的涌现现象？

- 为什么提示学习能奏效？这是否暗示“基础模型”内部可能自发地产生了若干功能分区，而一个个提示学习恰好提供了启用一个个功能分区的钥匙？

- 如果是这样，功能分区的分布可能是怎样的？由于“基础模型”的核心训练算法极其简单（语言模型或完形填空模型），这又隐含着什么深意？

我个人认为，对“基础模型”科学意义的探索也许大于其工程意义。如果其中确乎蕴涵着上述一二玄机，那么这将对人工智能模型的全新发展具有深刻的启迪性，“基础模型”也会出现“山重水复疑无路、柳暗花明又一村”的全新气象。此外对脑科学、认知神经科学研究也可能富有启发性。

第二个挑战：智深者为上?

这是人工智能的“初心”和永恒梦想，与第一个挑战的思路相去甚远，但其必要性毋庸置疑。这里举例说明。

前文提及的机器翻译先行者Yehoshua Bar-Hillel，1960年发表了一篇长文《语言自动翻译的现状》，对机器翻译的前景进行了展望。文中他举了一个对人来说易如反掌，但对机器翻译来说异常棘手的一个句子（注意其中的 The box was in the pen）：

> Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

其中pen有两个意思：“钢笔”和“围栏”。要正确地翻译成“围栏”, 机器需要明白介词in的意思，同时具备相关的世界知识。我们把这个简单的英文句子，送给用深层神经网络和大数据武装到牙齿的机器翻译系统。

> 谷歌翻译结果：盒子在笔里。
> 百度翻译结果：盒子在钢笔里。

60多年过去了，还是没搞定。

可喜的是，在“力大者为王”波澜壮阔、摧枯拉朽的大势下，一批学者仍在坚持并积极倡导小数据、富知识、因果推理等“智深者为上”的下一代人工智能发展理念。不过目前研究进展不大。这条道路上有两个难以逾越的“拦路虎”。

一是形式化常识库和世界知识库依然严重缺乏。Wikidata之类的知识图谱貌似规模庞大，但如果稍微审视一下就会发现，它所覆盖的知识范围仍然十分有限。事实上，Wikidata存在明显的构成性缺失，多是关于实体的静态属性知识，关于动作、行为、状态，以及事件逻辑关系的形式化描写则几乎没有。这就使得它的作用域严重受限，实际效能大打折扣。

二是系统性获取“动作、行为、状态，以及事件逻辑关系”之类形式化知识的能力依然严重缺失。对开放式文本（如 Wikipedia 文本）进行大规模句法语义分析是必由之路。但很可惜，目前这个句法语义能力还不太具备（虽然近年来借助深度学习方法，已经有了长足进步）。

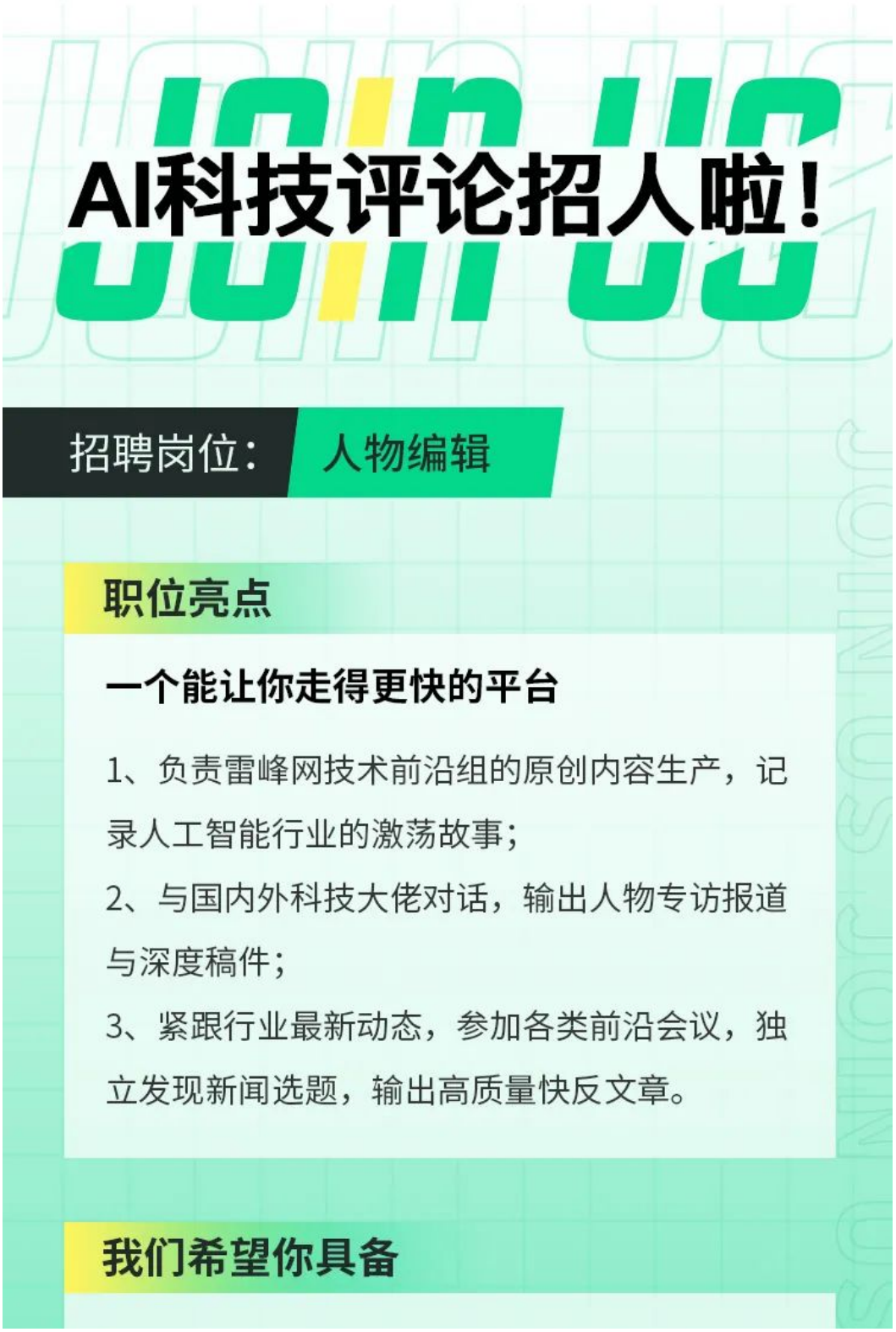
这两个“拦路虎”必须想办法解决。否则，巧妇难为无米之炊，这条路不易走通。

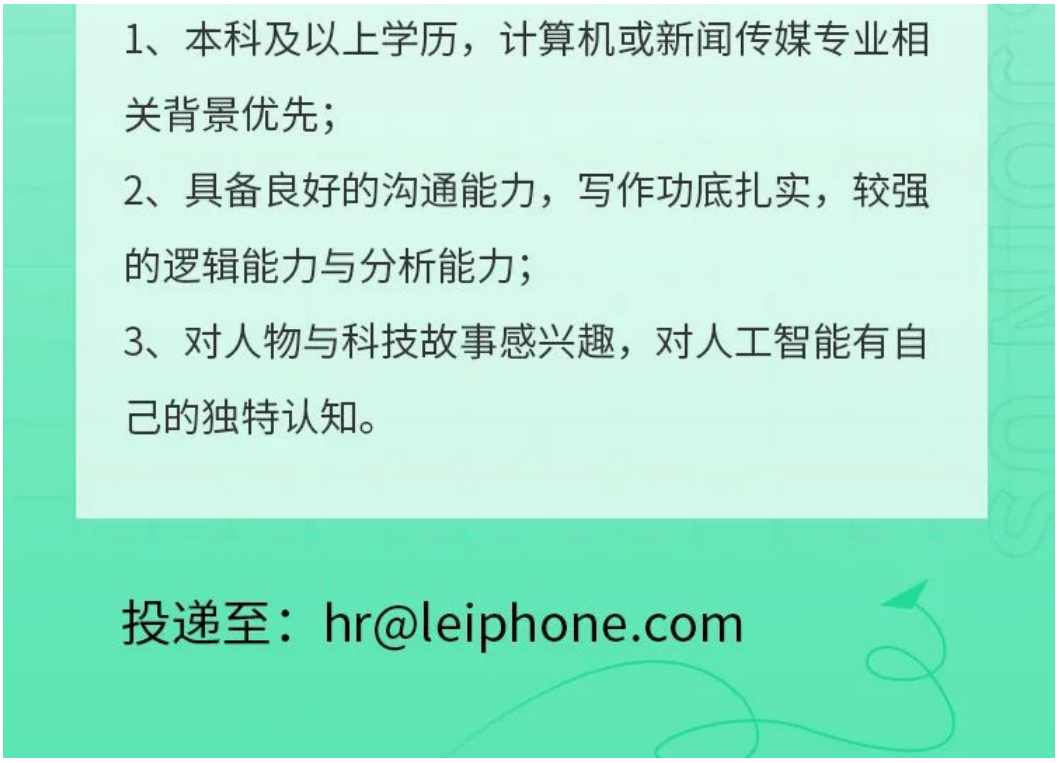
上述两大挑战，其实也是整个人工智能领域所必须面对的。

4

结束语

自然语言处理一路走来至今日，形成了“力大者为王”和“智深者为上”两条道路。前者道路宽广，顺风而下，但貌似快走到尽头；后者道路狭窄，逆风而上，但应该会悠长隽永。前看两者可以并行不悖，互相借鉴，互为支持，如“基础模型”可望有效提升大规模句法语义自动分析的能力，从而为大规模知识自动获取提供前提条件。“基础模型”可能包藏了某些深邃的计算机理或奥秘，或导致大的“峰回路转”，值得密切关注。未来 10 年，自然语言处理在研究和应用上整体上创造一个恢弘格局，并对人工智能领域的发展做出关键性贡献，是可以期待的。





喜欢此内容的人还喜欢

剑桥高级机器学习讲师Ferenc Huszár评马腾宇新作：它改变了我对上下文学习的思考方式
AI科技评论

计算机体系结构顶会 ASPLOS 2022 最佳论文出炉
AI科技评论

中国首次！清华刘奕群团队获得WSDM 2022唯一最佳论文奖，港中文获得「时间检验奖」
AI科技评论