MT to Death，专访 ACL Fellow刘群，一个NLPer的极致表白

原创 智源社区 智源社区 2022-03-14 14:17

收录于话题
#前沿进展 37 #机器翻译 3

导读：2022年1月6日，国际计算语言学学会ACL正式公布了2021年ACL Fellow名单，**机器翻译专家、华为诺亚方舟实验室语音语义首席科学家刘群**当选为全球八位新晋ACL Fellow之一。

我们了解到，刘群老师不仅是一个出色的科研人员，他还有另一个身份是一位微博大V，他的微博名称有一个鲜明而有趣的后缀**“MT to Death”**，这大概是他对MT（机器翻译）始终如一的表白。而“机器翻译”也是他当选2021 ACL Fellow的标签之一。

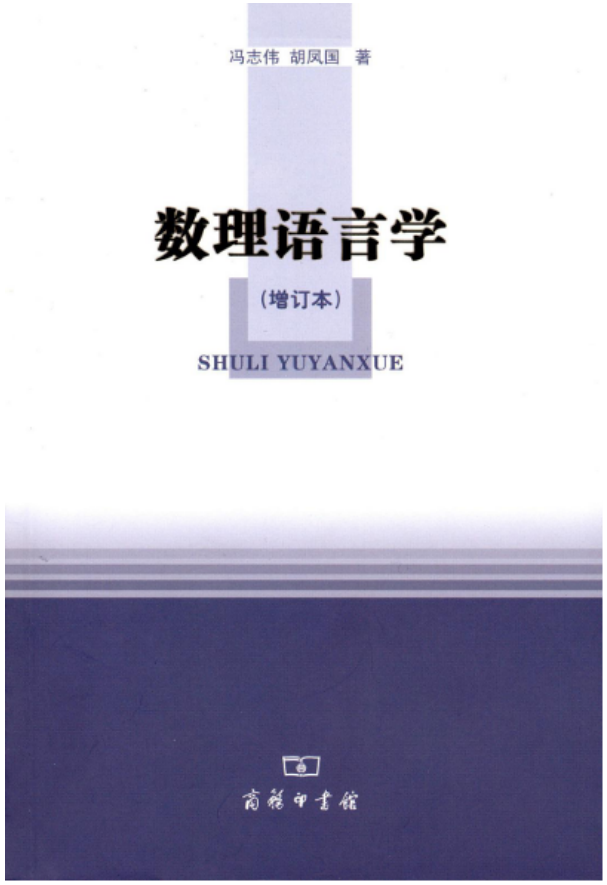
借此契机，我们对刘群老师进行了一次专访，就其个人跨学术界和工业界的研究经历，针对自然语言处理领域细分趋势及展望，对年轻科研人员的寄语等方面进行了一次深度访谈。

访谈对象：刘群
撰文：Lilian
编辑：梦佳

01

从《数理语言学》入门到真正开始从事机器翻译研究

1984年，刘群刚刚入学中国科学技术大学，攻读的是计算机科学技术专业，本科时期，在中科大合肥校园的图书馆，刘群第一次看到了《数理语言学》，这是我国计算语言学的开拓者之一、世界上第一个“汉语到多种外语机器翻译系统”的研制者冯志伟老师所著的一本书，书中系统地、全面地、深入浅出地介绍了代数语言学、统计语言学、应用数理语言学三个部分的基本知识。 **正是这本书在那个网络和传媒尚且不发达的时代，带领刘群认识了自然语言处理这一新的研究领域。**



冯志伟老师的《数理语言学》，上图为刘群当时看的早期版本，下图为后来的新版

而在自然语言处理当中，刘群最早接触到的是机器翻译。20世纪80年代末期，新的机器翻译系统大量涌现。1989年从中科大本科毕业，他被保送到了中科院计算所读硕士，开始参与一个英汉机器翻译项目的研究。**这也是他走上机器翻译这一方向的真正开端。**

02

从艰难起步到一步步走向成功

1992年，刘群硕士毕业，他留在计算所，1993年，**他在非常困难的情况下开始了独立的机器翻译研究。**刘群选择了汉英机器翻译这一难度更高、对汉语意义更大的课题。随后的研究工作中，他与北京大学计算语言学研究所俞士汶教授建立了长期深入的联系与合作。直到1998年，刘群团队和北大计算语言所联合开发的汉英翻译系统在863中文信息处理与智能人机接口技术评测中取得了较好的成绩。至此，他的机器翻译研究迎来了一个小的高潮。

1999年，刘群报考了北大的在职博士，被录取为俞士汶老师的在职博士研究生。1999年末，俞士汶老师得到一个973子课题“面向新闻领域的汉英机器翻译系统”，刘群以计算所员工和北大博士生的双重身份，担任这个课题组的技术负责人，继续从事机器翻译研究工作。

2004年，刘群在北京大学获得博士学位，并回到计算所继续从事机器翻译研究。2005年，他在计算所评上了研究员职称。从这以后，刘群开始以自己名义正式招收博士研究生，并组建了一支充满活力的研究团队。

在1990年代到2000年代初期，国际上软件开源运动正在兴起，但在学术界，开放源代码还没有形成风气，可获得的开源代码和开放数据资源都十分有限。在那个开源资源非常有限的年代，国内的研究者只能通过有限的学术刊物和会议论文了解国际上最新的研究动态，但这些最新的技术和方法的大部分实现细节，都隐藏在论文介绍的原理和公式背后。

刘群和他的团队为了掌握国际上最先进的技术，每次看到国际上有什么重要的研究进展，**都有一项“必然操作”，那就是从各个角度、用各种方案对这些方法进行还原实现 —— 看数据、清理数据、尝试各种技术路线、调试代码** …… 正是他们这种务实的“啃硬骨头”的做法使得他的研究团队较早掌握了当时国际上先进的统计方法，并在自然语言处理和机器翻译研究取得了一系列突破。**其中，中文分词系统ICTCLAS和基于《知网》（Hownet）的词汇语义相似度计算两项工作便是最好的证明。**

在当时国内缺乏开源代码的环境下，刘群和他的团队把中文分词系统进行了开源，这是当时性能最好的中文分词系统，也是当时唯一可公开获取的系统。另外，刘群还将他开发的基于《知网》（Hownet）的词语相似度工具的可执行代码公开出来提供免费下载，这两项公开的成果成为当年很多做中文自然语言处理同行所使用的最基础的工具，在国内产生了很大的影响。

2002年，刘群团队作为唯一来自中国的研究机构参加了美国NIST机器翻译评测，虽然首次参赛的结果让人大失所望，但他却深切感受到了统计机器翻译方法相对于传统基于规则的机器翻译方法的优势，并痛下决心从传统的规则方法彻底转向了统计机器翻译方法。

在接下来2005年的NIST评测中，他们取得了第五名，证明他们开发的机器翻译系统达到了国际同类研究机构的先进行列。同年，刘群的博士生刘洋第一次在 自然语言处理顶级会议ACL上发表了论文，紧接着他的博士生刘洋和熊德意再次在2006年ACL上发表了两篇论文，提出了两种新型的基于句法的统计机器翻译方法。**在这之前，国内研究机构在ACL上总共只发表过1篇论文，而刘群的团队在两年内连续发表了3篇论文**，这在当时国内的研究机构中是非常罕见的，引起了很多人的关注。

后来刘群的团队持续在基于句法的统计机器翻译方面做了一系列工作，这在国际上也产生了很大影响。因为经典的基于短语的统计机器翻译方法捕捉句法结构的能力比较差，对于类似中英文这种结构差异较大的语言之间的翻译质量影响很大，而刘群团队的工作主要是基于源语言端句法的统计机器翻译方法，在这一方向上发表了一系列论文，这些工作达到了国际先进水平。

03

两位院士的拷问：超越硬件的约束

这段时间，刘群团队研究工作已经在国内外有了一定的影响。由于统计方法需要使用大量的并行计算资源，刘群在计算所内给领导汇报工作的时候经常会说起计算资源的不足影响了研究的进展。有两次在不同的场合，**李国杰院士（也是当时的计算所所长）和高庆狮院士都向他提出过同一个思考题：没有机器的约束，给你无限多的资源，能把机器翻译做到什么程度？**

李国杰院士（左），高庆狮院士（右）

刘群说，这个问题给了他很大的震撼，迫使他更多更深入地思考机器翻译的长远发展问题。众所周知，由于摩尔定律的原因，机器的性能在快速提升翻倍，虽然如此，大部分研究人员在真正做研究的时候还是很受实际条件制约，经常会觉得机器不够用，并没有太多时间去深入思考更长远的问题。

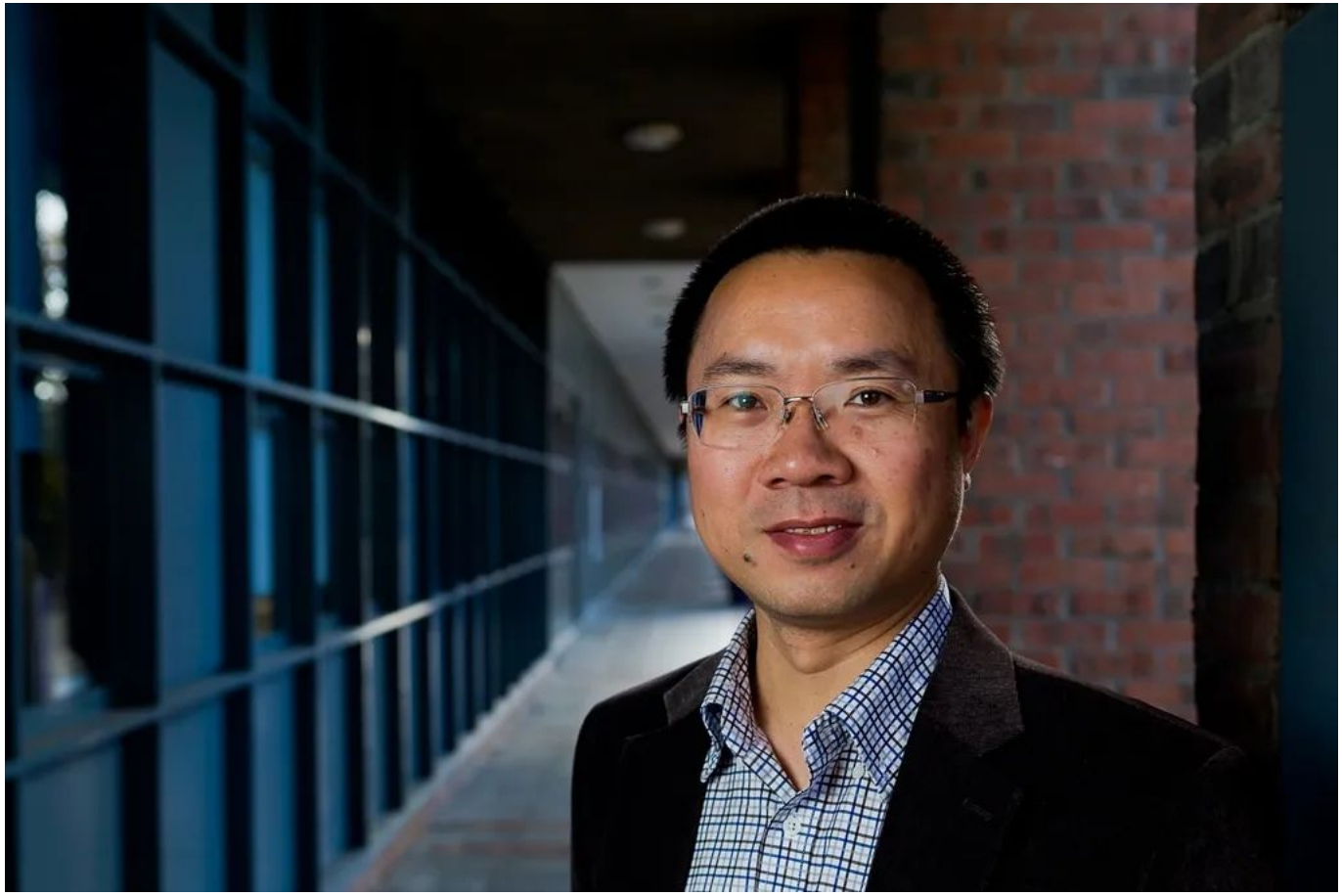
而两位做硬件体系结构出身的院士考虑问题的角度却不一样。他们更多考虑的是：其他研究领域，包括人工智能，会给硬件和体系结构带来怎样的挑战？而硬件和体系结构的改进，反过来又将如何促进其他领域的进步？两位院士提出的这一问题，刘群当时无法给出答案，但这个问题却极大地开拓了他的思路：**做研究的人想问题要超前一点，如果不完全考虑硬件约束的情况下，研究能走多远。**带着这样的问题去思考，做研究时考虑的角度和深度就跟原来完全不一样了。

04

就职都柏林城市大学：主动寻求转变

2012年7月，刘群加入了爱尔兰的都柏林城市大学担任教授职务，并在爱尔兰下一代本地化技术研究中心（简称CNGL，后改名ADAPT研究中心）担任机器翻译方向的主题负责人。当时的中心主任是来自德国的Josef van Genabith教授。

在爱尔兰六年从事科研和教学工作的经历，给刘群带来了很多新鲜的体验和感受。除了语言、生活、文化和科研体制的冲击，研究的内容和方法也在发生改变。



刘群在爱尔兰都柏林城市大学

刘群发现，虽然都是机器翻译研究，但CNGL/ADAPT中心关注的重点跟他自己原来的关注点非常不一样。刘群原来在计算所的研究，非常关注机器翻译的核心模型和方法，而CNGL/ADAPT这边更多关注解决机器翻译在实际应用中所面临的一些问题，比如**翻译记忆、术语翻译、翻译质量评估、译后编辑、交互翻译**等等。他慢慢意识到这些他原来所忽视的课题的研究价值，并开始带学

生在这些方向做出了一些有影响的工作。比如他指导他的学生Chris Hokamp完成的词汇约束的神经机器翻译解码方面的工作，就是在神经机器翻译框架下首次提出了一种给定术语约束解码的方法，这一工作被很多后来的研究者引用和改进。

在这一段时间，整个机器翻译领域也发生了一次重大的变革：**从统计方法转移到神经网络方法。**与以往的情形类似，这次变革也是由机器翻译外的技术进步带来的，一些深度学习研究者在语音、图像等领域取得巨大成功后，开始把目光瞄准了机器翻译，并取得了初步的成功。而很多原来的机器翻译领域的研究者，在这一变革来临的时候还有点犹豫观望，并没有意识到这个变革会给机器翻译领域带来颠覆性的影响。

刘群是原机器翻译领域研究者中较早主动拥抱这一变革的人之一。为了更好地在机器翻译领域推广这种先进的技术，刘群带领他的团队于2005年10月在都柏林城市大学组织了一次为期一周的DL4MT Winter School（机器翻译的深度学习方法冬季学校），邀请了三位这一领域的顶尖学者来详细讲解深度学习的理论和方法及其在机器翻译中的应用。这次活动取得了非常大的成功，吸引了来自世界各地的近百名研究者参加，对深度学习方法在机器翻译领域的传播和推广起到了非常积极的作用。不仅如此，刘群还让自己指导的博士生全部转向深度学习方法，并在这一领域做出了很多早期的探索性工作。

据他所说，在都柏林城市大学六年的研究工作对他来讲实际上带来了两个方面的转变，既是从统计机器翻译方法向神经网络翻译方法的转变，也是从理论模型研究向理论与应用并重的转变。作为一个中国土生土长的研究人员，这一段海外从事教学研究工作的经历对他来说也是一笔非常宝贵的财富。

05

加盟华为诺亚方舟实验室：开启新的篇章

2018年7月，刘群离开了工作26年之久的学术界，加入华为诺亚方舟实验室担任语音语义首席科学家，开始了他的职业生涯的新的篇章。

华为诺亚方舟实验室为刘群的研究工作提供了一个全新的更加广阔的平台。一方面，原来在高校面临的资金、人才和计算资源的缺乏等制约因素大大缓解，**另一方面，他也需要带领更大的团队，并需要面对论文发表和成果落地的双重挑战。**但刘群很快适应这个平台，并且渐入佳境。

刘群说，在研究工作中，他经常会有些大胆甚至天马行空的想法，原先在学术界，由于各方面资源的约束，这些想法通常只能是想想而已，而到了华为，很多原来看似不可能的想法，都有了尝试的机会，而得益于华为丰富的产品线，甚至原来一些看似没有太多实用性的想法，都有可能找到落地的场景并受到产品线的欢迎，这让刘群收获了非常大的成就感。

在技术落地方面，诺亚方舟实验室有比较成熟的管理和运行机制。刘群团队的工作涉及到和业务团队的灵活配合，“我们的工作是不跟某个产品绑定的，但要在公司内部证明我们对产品做贡献，**在需要的时候，我们会与业务团队做短期内的强绑定，在某一段时间内密切配合产品团队做好某项任务后，我们又会退出**”。

刘群谈到，在加入诺亚方舟后，恰逢预训练语言模型的兴起，刘群的团队很敏锐地抓住了这个机会，迅速投入大量资源开展研究，同时与产品团队合作，探索如何将预训练语言模型落地到产品中。由于预训练语言模型规模巨大，占用空间多，推理速度慢，模型的压缩和加速成为产品落地面临的关键问题。

为了解决这一问题，刘群团队在很短的时间内，**提出了一种基于知识蒸馏的预训练语言模型压缩加速方法：TinyBERT。**采用这种方法，模型大小可以压缩到原始BERT模型的1/7，而速度提高了9倍。而他们的团队与多个产品团队合作，很快将TinyBERT运用到手机、终端等设备上，使得华为成为世界上最早将预训练语言模型大规模应用到产品中的公司之一。

“这是我们最早取得的突破，目前华为各个产品线基本上全部都应用了TinyBERT的技术，对公司做出了重大贡献，我们也非常高兴。”TinyBERT不仅在应用中起到了非常好效果，在研究界也产生了很大的影响。TinyBERT的论文在EMNLP2020会议上发表后不久，很快就成为该次会议引用最高的论文，目前引用次数已经达到500多次，成为预训练语言模型压缩中的经典工作。

06

热点探讨：语言模型是否越大越好?

此次采访，在分享个人学术经历之余，刘群也就机器翻译和自然语言处理领域的重点研究话题表达了他的看法。这几年，随着GPT-3一类的超大规模语言模型的推出，**一场预训练模型参数竞赛也随之而来，是否参数越大越好?**

他认为，模型规模大本身没有太大的意义，并非越大越好，关键是规模扩大后能够带来突破，这个才是最为重要的，比如说，最早的预训练语言模型（如BERT）为我们带来的突破之一就是“预训练+微调”的模式，原来，我们要为每个问题（下游任务）设计单独的模型，而BERT推出以后，对于大部分NLP任务，我们都不再需要重新设计新的模型，直接采用BERT加上少量下游任务数据微调即可。而更大规模的GPT-3模型推出以后，又在零样本和少样本学习上取得了突破。对于一些全新的NLP任务，甚至可以不需要训练，或者只提供几个简短的例子就可以直接解决这个问题，这在以前是很难想象的。

另外，超大模型有可能为产业界带来了很大变化，具有超强能力的大模型的能力以后可以放在云端，供大家调用。目前，像GPT-3这样的超大模型只能部署在云端，这就形成了中心化的AI能力，而这种能力是一般的中小型语言模型所不具备的。这种超大模型的部署，一般的小公司或科研机构也无法承担，这就需要在模型的部署和应用方式上进行改变。

超大模型以后也需要进行压缩，但由于模型太大，压缩整个模型是不现实的，而是应该根据特定任务的需要，抽取其中某些部分，蒸馏到小模型中。**这就需要开发新的模型压缩加速技术。**

有人认为，如果要把大模型压缩成小模型来用，为什么不直接训练小模型呢？刘群对此表示，用小模型直接训练得到的效果，通常都比大模型压缩以后得到的小模型效果差，因为大模型压缩后得到的小模型可以继承大模型的丰富的知识，这是直接训练小模型无法得到的。所以大模型在应用中仍然具有明显的优势，但如何发挥大规模模型的优势，这中间还需要做很多研究。未来，模型发

展的趋势应该是各种大小的模型互相协作，协作的模式可以千变万化，以应对各种不同的应用场景。在有些场景可能需要把大模型压缩成小模型，而有些场景则需要云边端协作。又比如，我们还研制了一种自适应大小的DynaBERT模型，可以方便地对它进行裁剪以满足不同场景的应用需要。

07

NLP未来展望

谈到NLP领域的细分及未来展望，预训练语言模型的出现为自然语言处理带来了新的研究范式。除此之外，去年以来，跨模态让大模型在视觉上带来了惊喜。对此，刘群表示，未来非常期待语言模型在**知识处理、常识处理**方面实现更多突破，也希望看到更多多模态应用带来的惊喜。

谈到机器翻译研究，刘群表示，虽然文本翻译目前取得了很大的成功，**但实时语音翻译或自动同声传译目前还面临着很大的挑战。**如果说文本机器翻译目前能够满足大部分场景需求，那么实时翻译还处在起步阶段。但挑战越大，研究蕴含的乐趣就越多，他认为相比其他的研究方向，实时翻译是一个研究起来非常有意思的领域。此外，目前篇章翻译也还存在很多问题，如论文、小说的翻译，最大的问题就是术语前后不一致，在这方面解决方案之一是引入符号推理，不仅可以提高模型的可理解性，在减少翻译一致性错误方面也具有较好的前景。

谈到对话系统，刘群认为，对话系统的研究难度要高于机器翻译。对话的生成缺乏源语言语义的约束，而涉及到的问题复杂程度是没有任何限制的。在闲聊对话方面，用大模型生成自然的响应目前在自然性已经可以做得比较好，但在实际应用中，也还面临很多问题。比如研究人员需要在这一基础上对系统进行适应性调整，目的是保证安全性，避免出现消极或者冒犯性质的语言，还需要避免出现偏见或者歧视性内容。

对于任务型对话，简单对话如订票、订酒店等已经做得很好了。但对于复杂的业务场景，比如移动公司的客服，它们有上百种产品，在这种情况下定制一个很好的对话客服系统难度就很大。通常，研究人员会收集对话语料，来训练一个对话系统，但对于带有复杂逻辑的业务系统，这种做法是远远不够的，在这种情况下如何快速搭建一个好的对话系统目前还没有很成熟的办法，这也将是今后值得研究的一个重要方向。

谈到问答系统，一个难点是开放式问答，由于涉及的范围没有任何限制，开放式回答通常要利用检索到的多个文本进行推理并生成答案。

对话和问答都涉及自然语言生成技术，这是NLP中比较难的研究方向，也是今后的研究重点。自然语言生成的另外一个问题是hallucination，指模型会胡说八道，意即生成一些毫无依据的内容，**如何解决hallucination问题也会是自然语言生成今后的重点研究方向。**

08

一个NLPer的科研信条：教学是一生的事业

当被问及最引以为豪的是什么时？刘群老师的回答很简单：学生。

采访中，我们能清晰地感受到刘群老师提及学生时的自豪和欣慰之情。**他认为，培养出了一批热爱机器翻译并至今一直从事机器翻译研究和开发的学生，这就是他心中分量极重、甚至远超其科研成就的一件事。**

这些学生中，有些已经是活跃在高校和科研机构中科研人员，如冯洋（计算所）、刘洋（清华）、熊德意（天津大学）、苏劲松（厦门大学）、侯宏旭（内蒙古大学）等，已经成为我国机器翻译领域青年研究人员中的佼佼者，更多的学生则进入了企业界。特别值得一提的是，国内一些主要的大型IT企业的机器翻译团队的负责人或者核心技术人员，几乎都有刘群的学生，如百度、有道、腾讯、阿里、小米、字节跳动、华为等等，这些学生取得的每一个成就都让刘群感到发自内心的高兴和骄傲。

刘群团队获ACL2019最佳长论文奖，其中刘群的学生冯洋（计算所）为论文主要指导者

持之以恒是刘群老师一直坚信和实践的科研信条。谈到对学生和年轻学者的建议，他表示，学生要清楚自己的长处，做到这一点虽然不容易但这对科研是至关重要的；对于年轻学者，刘群老师建议，一定要建立自己的学术标签，也就是一以贯之的研究主线，**可以换工作、换单位、换课题，但一定要坚持一个主线，这样才能形成长期影响力。**

此外，刘群老师也建议从事AI研究的青年科研人员在规划自己的学术生涯的时候，不管是长期来说有志于在高校从事科研教学工作，还是希望进入企业从事应用研究，都应该找机会进入类似华为这样的大企业AI研究部门工作一段时间，积累一些工作经验，因为AI是一个有非常强烈的应用背景的研究领域，在大企业能有机会接触到各种AI的真实应用场景，有助于研究人员深入的理解AI问题，这对于他们今后的研究工作是非常有助益的。

刘群的微博里有一个高频tag #自然语言理解太难了#，所发内容大多是让人啼笑皆非的“理解谬误”，连人类自己在自然语言理解上都会错误百出，更何况机器呢？是的，**Machine Translation To Death**，机器翻译的漫漫长路，还需要艰难跋涉，而始终如一的坚持才能收获至宝。

-END-