除了深度学习，机器翻译还需要啥？

眼球不够，八卦来凑

以一个“八卦”作为开头吧。

本文开始要写作的时候，翻译圈里出了一个“爆炸性”的事件。6月27日下午，一个同传译员在朋友圈里爆料：某AI公司请这位译员去“扮演”机器同传，制造人工智能取代人工同传的“震撼”效果。

这个事件瞬间在译员群体的朋友圈、微博、微信群引爆了隐忍已久的火药桶。因为过去几个月来，隔三差五就冒出一个号称要取代同声传译的翻译机，尤其是一篇题为《刚刚宣告:同声传译即将消亡!》的微信文章，在六月下旬铺天盖地的充满了一堆有关的或者无关的公众号，不知道带来了几个10万+。几乎每个翻译行业的从业者，都收到了朋友略带

同情关切的口气转过来的那篇文章，从一开始的“呵呵”到逐渐不堪其扰，终于被上面这个事情彻底激怒了。微博上的@同声翻译樱桃羊 呛声道：

> 还有台上的演讲嘉宾，光鲜亮丽的，德高望重的，一边享受着我们的服务，一边儿说以后同传都要失业，骨子里透着对这个行业的轻慢——“你们以后都是要被替代的工种，我们何必在乎你们的感受？”有本事不要请我们，既然请了我们，在说这句话之前，能不能跟现场辛苦工作的同传说声抱歉？

“积怨”深到了什么样的程度，可见一斑。

这件事后来有了一个略带戏剧性的转折。两天后，《消亡》文中被吹捧的晓译翻译机的制造者，科大讯飞（注意：没有证据表明上面的造假和讯飞有关）在其公众号上发布了一篇题为《拒绝神化 人工智能技术需踏实前行》的文章。重点在这两段：

> 目前，机器翻译已经取得非常大的进步，在衣食住行等常用生活用语上的中英翻译可以达到大学六级的水平，能够帮助人们在一些场景处理语言交流的问题，但距离会议同传以及高水平翻译所讲究的“信、达、雅”还存在很大的差距。
>
> 讯飞一直所努力的，是希望通过语音转写和翻译技术帮助同传提高工作效率、减少失误，形成人机耦合的同传新模式，并不是去替代同声传译。

好吧，原来是“被神化”，原来“我是拒绝的”。但是不管怎么样，此文一出，又收获了不少不明真相群众的交口称赞。而义愤填膺的译员们，也迅速恢复了低调内敛的幕后工作者姿态。甚至连揭发造假者的呼吁，也立马无声无息了。

作为一个混迹在人工翻译圈的机器翻译工作者，我感觉到，这个事情暴露出了一些对翻译行业的很深的误解，无论是普通大众还是机器翻译/人工智能工作者。本文的主题，初衷就是探讨人机结合对于机器翻译发展的重要性。所以，有必要首先澄清一下这些误解。GitChat的读者们可能更期待技术深度，一开始我也打算写得更技术流一些。但是，上面这个事件出现之后，我觉得在关心人机结合模式的技术实现之前，我们还是得多探究一下问题的本质。雄心勃勃要替代人工的同行们，也得先知道要替代的到底是神马样的存在，不是么？

误解一：翻译就是简单机械的语言转换

对翻译工作的误解非常多，这点是最核心的，很多其他误解（比如会外语就能做翻译、翻译人才遍地都是、翻译不需要理解专业就能做等）和由此产生的轻慢（翻译没有技术含量、随便可以替代、不值钱等），正是根源于此。

为什么这么多人“一致”相信翻译即将被替代？除了博眼球的自媒体之外，“替代党”包括了太多的互联网大佬，比如库兹韦尔、李彦宏、李开复、王小川等等。我们来看看这背后的逻辑是什么。

李开复在其新书《人工智能》中提出了一个“五秒钟准则”：一项本来由人从事的工作，如果人可以在5秒以内对工作中需要的思考和决策的问题做出相应的决定，那么，这项工作就有非常大的可能被人工智能技术全部或部分取代。

然后基于“五秒钟准则”，该书预测：从事翻译、新闻报道、助理、保安、销售、客服、交易、会计、司机、家政等工作的人，未来10年将有约90%被人工智能全部或者部分取代。

显然，翻译不幸被首当其冲的归入“五秒钟”可以解决的问题。大概是同声传译这个带着光环的工种给大家带来了一个幻觉：翻译不就是几秒钟就出来的吗？而且，据说同传还是所有翻译工作里最难最贵的，那其他翻译岂不是更不在话下？

事实上呢？口译员们知道，为了准备一场会议口译，事先要提前做多少天的功课，会前要做多少沟通协调，会中要多注意随机应变；更不用说，达到可以做同传的水准，要经过多少年鬼知道经历什么的刻苦训练。用“台上一分钟，台下十年功”来形容口译员的工作，再贴切不过。而笔译的工作，也一点都不轻松，也同样需要多年的磨练才能产出合格的译文。

进一步说，这个误解实际上包含两个论断：

1. 翻译只是语言转换。
2. 语言转换是简单机械的。

实际上，这两个论断都是不成立的。

对于第一个问题，我们要追问一下翻译的本质。翻译是一种语言服务。对于语言服务，近期看到广东外语外贸大学的李瑞林教授给出的定义，我认为最接近其本质：语言服务是以语言资源为基础，以致知、赋能、移情为目标，实现知识和经验人际或组织间转移的社会经济过程（见《语言服务概念框架的再反思：存在依据、普遍本质及实践逻辑》）。可见，语言是翻译这种服务实现的一个载体，而不是服务本身。翻译工作带来的知识、经验和情感，才是最关键的东西。而这些东西，相信开复们都不会认为是机器很容易处理的（记忆性的静态知识除外）。

对于第二个问题，则要进一步追问“语言”的本质。我们知道，语言是信息的载体。比如我们要传达“苹果”的信息给另一个人的时候，不用非得拉着他去水果店或者某高科技体验店去才行。但是这个信息载体并不是无损的，我们用“苹果”这个概念，显然无法把具体的形状、颜色、触感、气味、效用等信息都全部传输过去，得靠对方把其他信息“脑补”出来。因此，语言实际上只是信息处理过程中的一个经过编码了的“快捷方式”。能否把快捷方式所代表的信息解码出来，对信息接收者的认知结构是有要求的。这就是所谓的“一千个读者就有一千个哈姆雷特”。一个翻译工作者既要做解码者，又要做编码者，必须在短时间内使自己的认知结构接近原文作者的预期，又要考虑到另外一种语言的读者的认知结构的差异。转换的难度可想而知。

因此，翻译这个事情，不是想当然的那么简单。在本文的预告贴里，我列举的几个语言特性中，“开放性”、“歧义性”、“演化性”等问题，给翻译带来了极大的困难。我们想要让机器翻译达到或者超过人工翻译，首先得正视其困难和价值，而不是靠将对方“简化”为

某种形式的机器。否则的话，我们岂不就相当于靠把国乒搞垮来实现“让国足达到国乒的水准”？

误解二：人工翻译就是“好翻译”

这点可能是人工智能工作者普遍的误解，但也是目前大家对机器翻译普遍非常乐观的一个原因。受“图灵测试”思想的影响，我们会把“让机器翻译给出好的译文”这个问题转化为“如果机器翻译给出的译文，人无法分辨是人还是机器做的，就是好的译文”。所以绝大部分的机器翻译训练，无论是统计机器翻译还是人工神经网络，都以和人工译文语料库的“最大似然度”为训练目标。也就是，想办法让机器译文看起来和平行语料的对译关系最接近。

这个假设又包含以下几个子假设：

1. 人工翻译水平是质量刻度线上的一个黄金分界点。
2. 我们可以从平行语料（只要足够多）中学习到这个分界点。
3. 普通双语人士可以很确定的分辨出译文质量是否过了这个分界点。

很遗憾，这几个假设，也是我们为了让问题有更良好的形式化定义，以及有相对一致的评价标准，而做出的简化假设。在我们离问题的“完美”解决方案还很远的时候，这些假设对于我们做出切实有用的近似解决方案，是非常有帮助，也是必须的。但是如果把在这些假设下得到的局部最优解的大幅进步等同于非常接近全局最优解了，就会产生即将冲破临界点的幻觉。这对于探求真正的真理，是不利的。

我们逐个说一下上面几个子假设。

第一，人工翻译水平显然不是质量刻度线上的一个点，而是上下界离得很远的一个区间，比如从30分到99.99分。其下界低于机器翻译的水准，是最正常不过的事情了。更要命的是，这个刻度线上的刻度值（如果以所需投入的努力作为间距衡量单位）并不是等距离分布的。打个比方，59分到60分如果间隔1厘米的话，95分到96分的间隔可能是1米，而98分到99分可能是几公里……所以，试问我们应该把哪个位置设为人工翻译的水平线呢？这里面学问可就大了——比如把分界点设在60分，让机器翻译从30分提高到57分，是不是可以说目标达成90%了？即使我们设定了99分的高标准，机器翻译从39分提高了30分到69分，离99分还有30分的时候，我们能宣称走完一半的路了吗？恐怕万里长征只是第一步。可是我们很容易想当然的这么来炫耀我们走过的路。比如去年谷歌GNMT的那篇论文，就是这样计算出来提升了87%（某个语种方向上，大家可以去找来原论文分析一下其评测数据和结论之间的关系），已经可以看到胜利在凯旋门下招手了……

第二，相比别的一些自然语言处理任务，机器翻译显得更成功，就是因为有平行语料这种天然带标的数据资源。只要平行语料是人工翻译的，我们似乎就可以将其作为衡量质量的黄金标准。然而，问题在于：

- 语料是有限的，但语言是开放的，和语料不匹配的，不代表是“不好的”。

- 语料是静态的，但语言是动态演化的——过去好的，现在未必好。在一个领域好的，换个领域未必好。对一部分人或场景好的，对其他人或场景未必好。

- 语料本身的质量可能是参差不齐的（因为人工翻译的质量是参差不齐的，参考上述第一条，暂且不说很多语料的来源也是机器翻译），尤其是海量规模的时候，也就是说，近似的也未必是好的。

因此，基于双语语料的质量标准，也是没有更好办法的办法，“黄金”度还是不够高的。

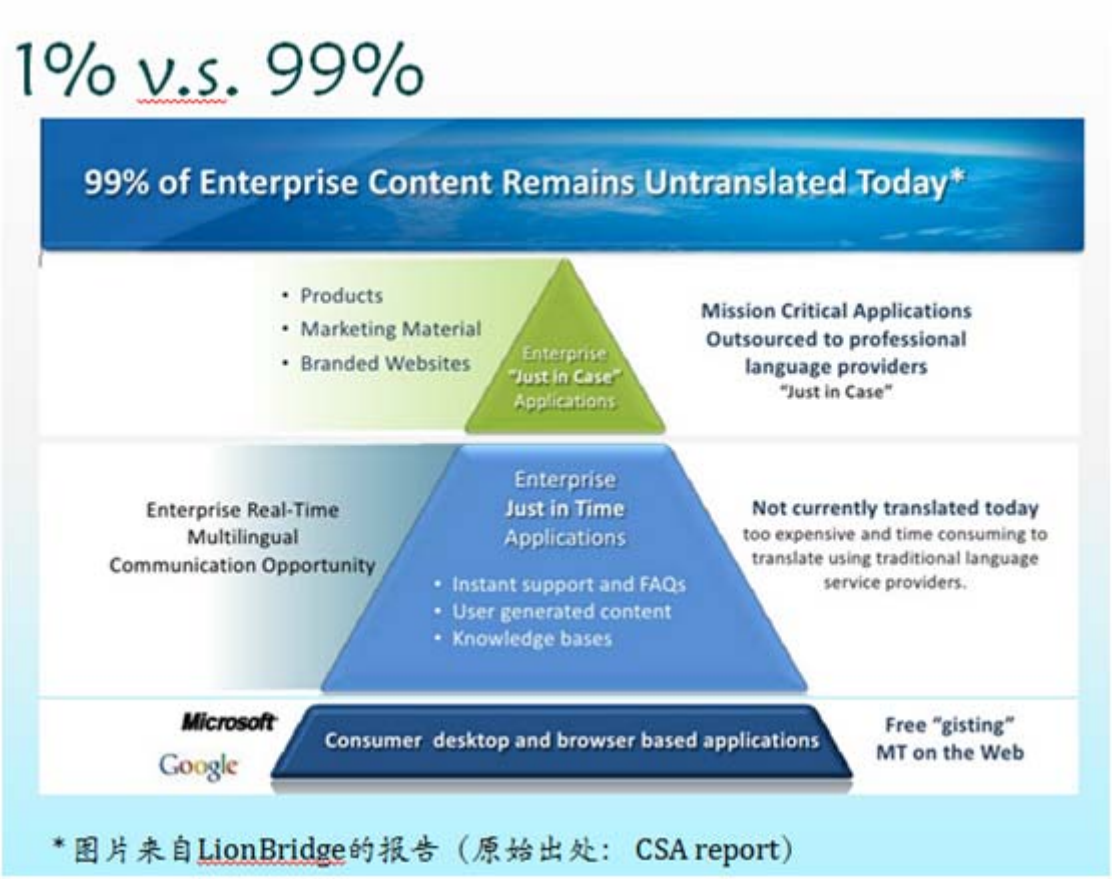
第三，假设我们确实可以找到足够好的人工译文作为质量标准，那么是不是任何一个双语人士都有足够的判别能力，来正确区分人工译文和机器译文呢？也就是说，会不会出现这种情况，一个机器译文确实是有瑕疵的，但某些人就是看不出来它与参考译文不一致的地方到底是好还是不好？答案是肯定的。这就像机器写诗，普通人的鉴赏能力或阅历经验有限，可能已经分辨不出它是不是机器写的，所以才有微软的机器人小冰潜伏在各个文学社区也没被发现。但是对文字敏锐的人，还是可以筛选出来哪些是好的诗歌——小冰出版的诗集，其实也是经过人工“精选”的。既然如此，不同资历、不同专业、不同文化、不同目标、不同条件的人，对翻译质量的认知也是不一样的。找什么样的人来作为图灵测试的鉴别者，也是一个需要仔细考量的事情。

三个子假设都与现实问题存在一定的差距，可见，我们目前所以为据的评价体系，还不足以带来足够充分的反馈。这点正是和下围棋的AlphaGo的最大不同。我们知道，AlphaGo最强大的地方，就是可以通过自己和自己下棋来不断提升水平，而自我对弈的前提，是明确的胜负判别标准。而翻译孰好孰坏，还没有很好的评价机制。也正因为如此，今年火的一塌糊涂的生成对抗网络（GAN）在机器翻译中虽然验证有效，但效果并不太惊艳。要想取得突破，在评价机制上必须多花点心思了。

误解三：翻译市场就那么大，机器做得多了，人的饭碗就小了

的确，我们如果把翻译市场比作围棋棋盘，人工翻译执黑，机器翻译执白，双方你死我活、拼命厮杀、攻城掠地、此消彼长，那就会存在所谓的“替代”问题。

可是，现实的翻译市场格局是怎样的呢？我们来看下面这张图。



这张图出自行业研究报告。如果把企业中的文字内容比作一个金字塔：

- 顶端的部分，是目前由人工翻译来完成的，包括产品资料、营销文案、品牌形象、法务合同等。
- 底端的部分，是目前由机器翻译来完成的，大多数是由用户发起的公开网页浏览等。
- 中间的部分，比如实时支持和FAQ、用户生成内容、企业内部知识库等，其实是没有被翻译的。

而这些没被翻译的内容，竟然占到了99%！

也就是说，无论对于人工翻译还是机器翻译，都有大片的处女地等待开垦，而且其中很大比例，可能需要人和机器携起手来，才能够给出可行的解决方案。

君不见，如今机器翻译用得最多的那些场景，比如旅游、电商、聊天社交，之前也并不是人工翻译的菜。而未来在客服、知识库、UGC等场景下的语言支持，基本上都要求既要有机器的快捷，又要有人工的可靠性或温度，一定是人机结合才能做到的。

更何况，上面的这个金字塔的体量也不是一成不变的，随着互联网和人工智能的发展，全球信息加速流动，内容规模不断膨胀，整体需求只会不断扩大，试问人机双方何时才能够在楚河汉界上兵戎相见？

进入正题

等等，预订时说好的内容呢？为什么总是在讲这些误解？

实际上，关于翻译行业或职业的误解还有很多，为什么就挑这三点来讲？不单是因为这三点最要害，更是因为澄清了这三点，我们的正题及预定通告中的第一问（为什么需要人机结合）的答案就非常清楚了。

1. 机器翻译要替代人工翻译，还有很远的路要走。

2. 人工和机器相结合，才能给出更好的解决方案，释放出原先被压抑的更大需求。

3. 翻译是知识、经验和情感的转移，人是实现这种转移的主体，现阶段只有通过人，才能更好的获取机器翻译所需要的知识和数据。

4. 缺少好的评价机制，很快将成为制约机器翻译进步的关键瓶颈，而语言的特性决定了，评价反馈不能来自语言本身，而只能来自于语言使用场景中的人。

对于致力于机器翻译/人工智能的同行，希望这篇文章，可以引起大家的一些思考。上述观点不一定正确。但是在密切跟进深度学习最新成果的今天，也许我们也应该适当低头想想，我们还缺什么，在可见的成果收割之后，还可以做些什么。

翻译圈的朋友，看到机器翻译还代替不了人工翻译的时候，是不是了松一口气？然而，这并不意味着机器翻译不会使现有的译员失业。再下一篇文章里，我将回答预定通告中的第二问（人机结合翻译怎么做）。顺带说说，机器翻译让译员失业的N种可能性。