注册 登录 投搞

首页 榜单 最新 视频 图集







科技 >> 信息

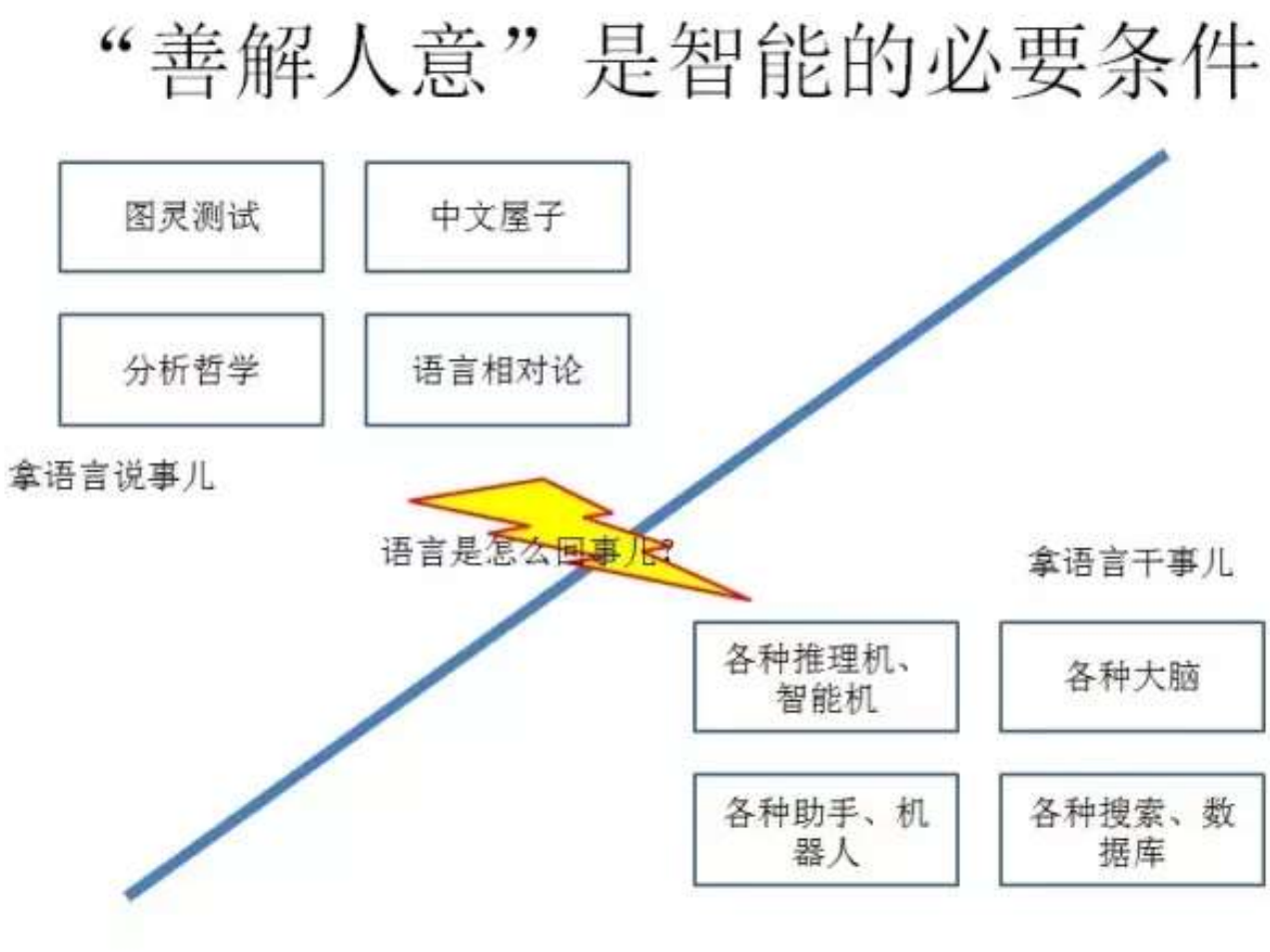
白硕：自然语言处理与人工智能

白硕 | | 2015-04-01

说起实现各种各样的智能系统，大家都认为“善解人意”应该是一个最起码的条件，一个绕不过去的条件。实现它的一个重要的桥梁就是语言。

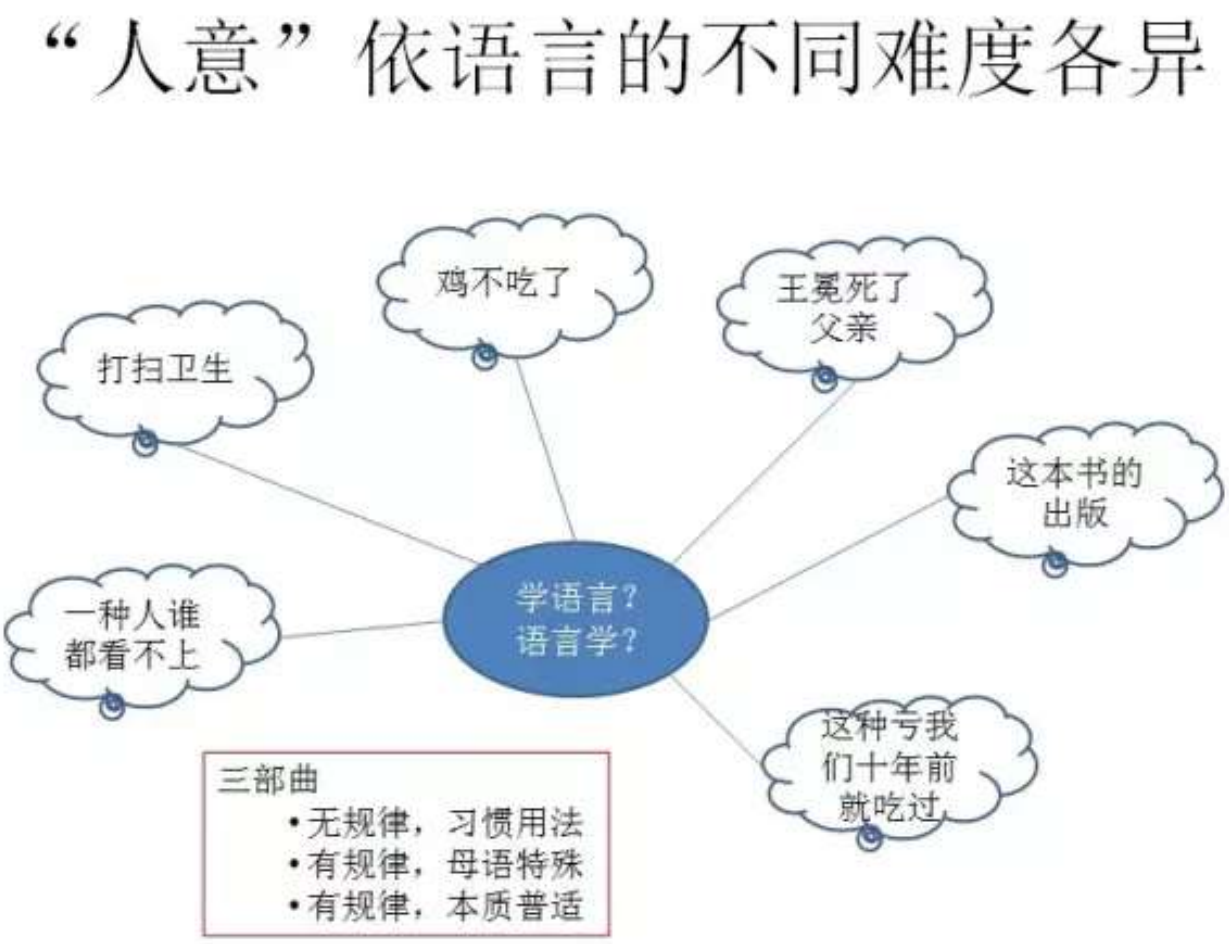
拿语言说事，这个我们见得很多了。刚才讲的图灵测试，实际上就是要通过这个对话，通过语言来判断隐藏在那个不可见的位置跟你对话的到底是人还是机器。

还有中文屋子，大家如果关心智能问题的话也都会比较熟悉，一个屋子里的人不懂中文，只懂英文，但可以执行特定的规则，把符号搬来搬去，最后效果是把英文翻译成中文，问题是，到底谁懂中文？



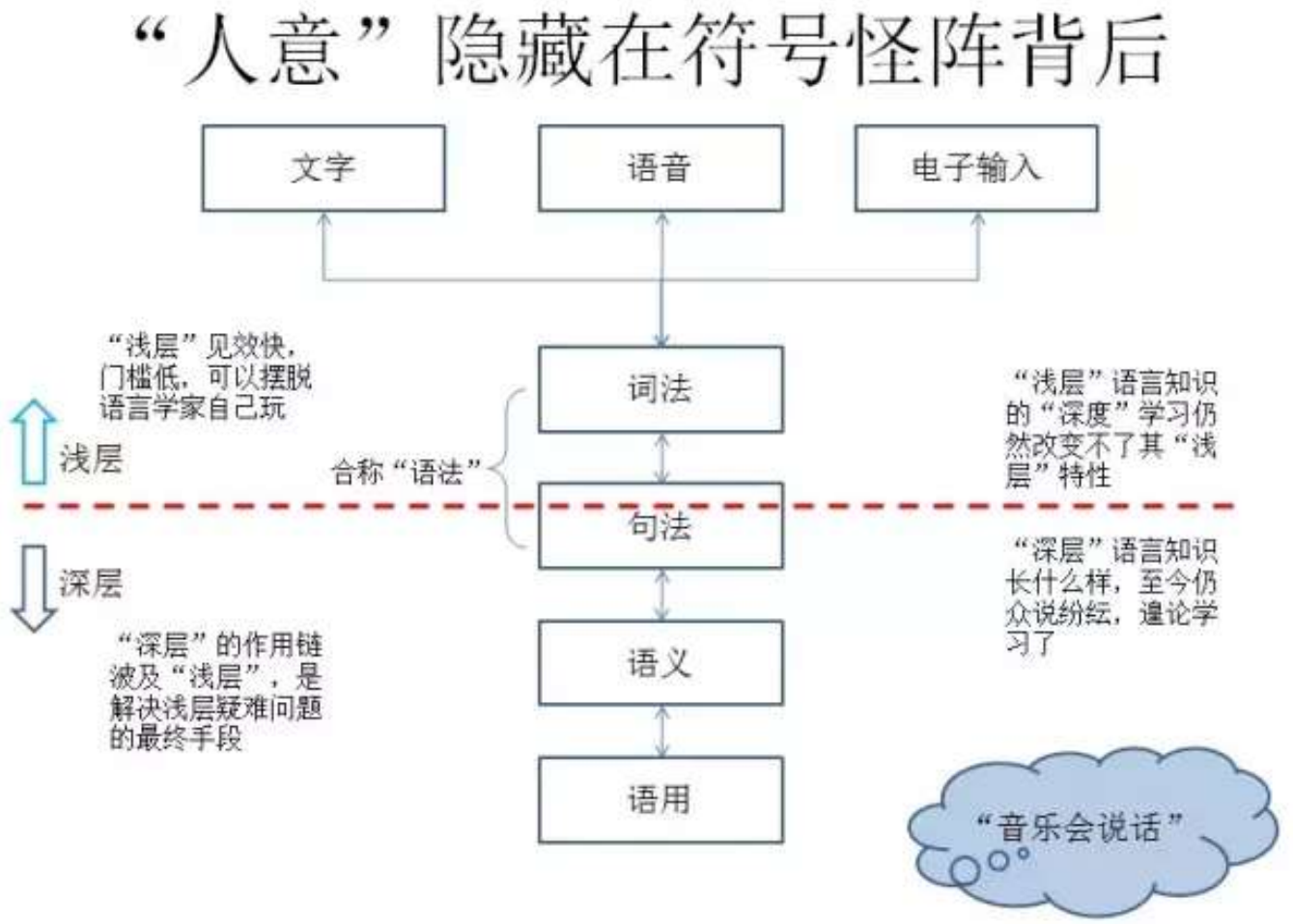
另外，从亚里士多德到布尔，从莱布尼兹到维特根斯坦，实际上哲学在20世纪初有一次重要的“语言学转向”，就是从拍脑袋的思辨到针对文本，看我们的理解，我们的定义，什么时候出现了问题。

语言学里面也有人思考，表现在外部语言和脑子里面的概念的关系，到底是不是一一对应的。今天看到杨义先老师讲“孝”这个字的英文到现在没有对应的贴切翻译。有人举例子说，蒙古人关于马的词汇好几十种，爱斯基摩人对雪的词汇也有好几十种，我们不生活在马和雪的世界的人就很少。是语言造就了概念，还是概念造就了语言？是全人类有共同的概念，还是有一种语言就塑造一套概念？看来这确实是个问题。



从干活的角度，我们看到了各种“大脑”计划，各种助手和机器人推出来，有各种搜索和数据库在精准化，拿语言去搜，也有各种的过去就有推理机，智能机，一直在延续着七八十年代的梦想，虽然进展不大。

但是所有拿语言说事儿的、用语言干事儿的，这中间都绕不开一个问题，即语言到底是怎么回事？这是我们真正搞自然语言理解，搞自然语言处理的人必须面对的问题。我们说要善解人意，人意在哪？它藏在符号怪阵的背后。我们来看这张图：



上面最表层的，可能是通过文字的方式表现的，可能是通过语音的方式表现的，也可能是通过电子输入的方式表现的。再向下有词法、句法和语义，乃至语用。这是语言学界标准的层次的划分，我们现在看中间的虚线，上方涉及到一点点的句法，句法的大部分在下方。以虚线为界，虚线之上的通常认为是浅层，虚线之下的认为是深层。请注意这里讲的是语言学知识的浅层和深层。浅层的这种语言模型见效很快，门槛很低，基本上可以摆脱语言学家，工程师自己可以玩，有数据就可以玩，或者是标注，找一些中专生就可以玩。然而深层的语言学知识，有一些地方可能说不清楚，可能众说纷纭，但是浅层解决不了的问题，最终要通过深层来解决。上海电台里有一档节目叫“音乐会说话”，不论是按照什么分词标准，很多的分词系统都会把它切成“音乐会/说话”，就是说“音乐会”当做一个词的切法是应该占优势的，但是恰恰这档节目说的是“音乐/会/说话”，它用到了深层的句法和语义，甚至到语用才确定是一个隐喻，只有到了这个层次反作用回去才发现正确的分词方案是什么。

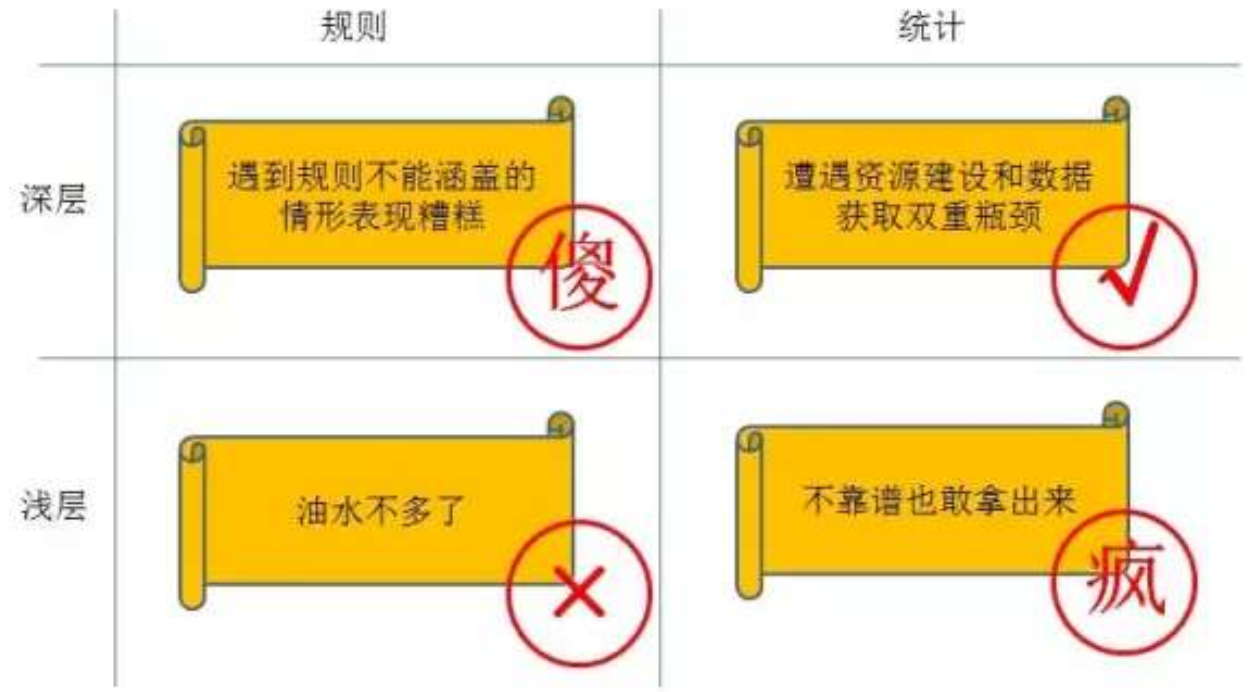
我们也看到了大家对深度学习充满了期待。但是必须说明，学习的度深和浅，跟学习对象在语言层次上的浅和深是两个概念。如果是在模型停留在浅层，就是把深度学习用到极致也有局限。

但是深层这个东西就复杂了，因为咱们讲证据，浅层的东西大家都看得见，深层有很多是看不见的，看不见就各说各的话，就没有一个统一的标准，所以这个东西就比较难搞，这是我们看到的一个现状。

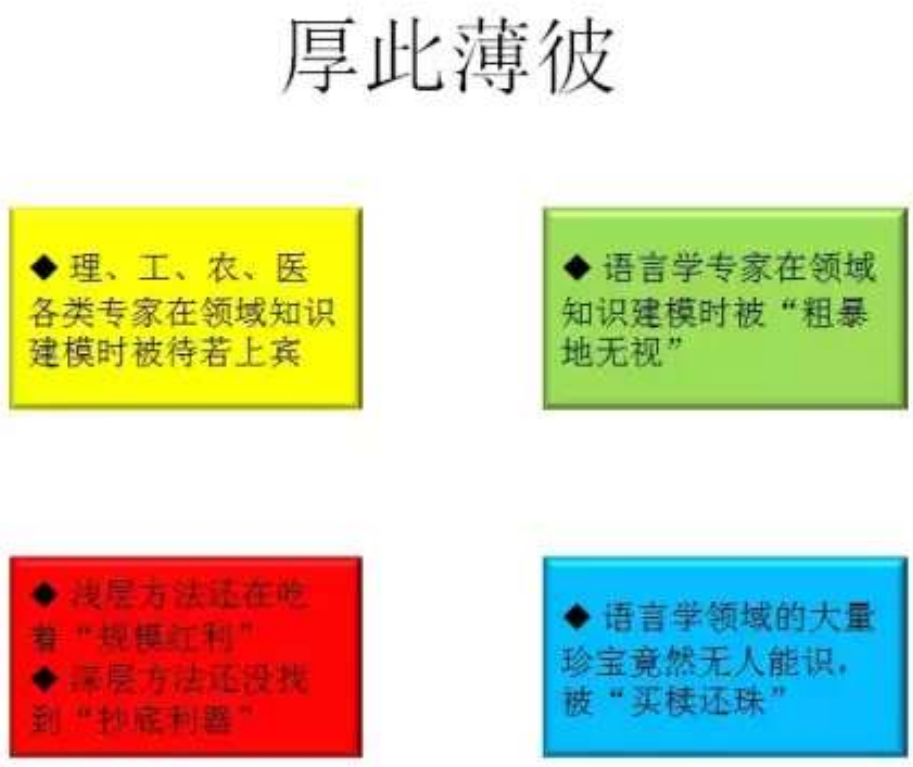
善解人意的“人意”好不好解，是不是跟语种有关系呢？在这张图上，我们看到很多中文难以处理的例子。很多评论的第一反应是中文博大精深，中文不一样，英语很好的方法到中文就不行。有的人就说我们比别人更懂中文。即使是知识图谱也出了个“中文知识图谱”，好象这个知识图谱到底是知识层面的还是语言层面的这件事情，都有了商量了。这说明了什么？说明我们还停留在我们对母语的认识的初级阶段。我们对母语的认识，和对人类共同的语言机制的认识，还没有统一起来，中间还有巨大的鸿沟。

一般来说，我们对母语的认识分成三个阶段：第一阶段，碰见这些例子，都认为这是习惯用法，根本没有规律的，我们也不关心这个，我们的指标很好看。这就完了吗？可是实际的问题并没有解决。第二个阶段，规律是有的，但我们的母语就是特殊，我们只能使用特殊的东西把这些问题解决了。是有这么一个见招拆招的阶段，但是作用很有限。真正的能够进到最后一个阶段的人很少很少。这种人既熟悉了母语的特殊性，同时又知道这个世界上关于语言共性普适的进展，而且还能把二者融会贯通结合起来。这种人太少了，而且这种人跟我们的交流太少了，所以我们不知道他们的存在，更不知道去挖掘他们的宝藏。

经常有一种说法，说基于规则的系统是傻子，基于统计的系统是疯子。基于规则的系统，在规则不能覆盖的地方表现极其糟糕。基于统计的系统，因为很多模型现在只能做到浅层，深层的不知道怎么做，所以实现也只能做到浅层，有局限，又不知道自己的局限在哪，就出现了明明不知道的，没有自知之明了，大家一看就知道是笑话，但是他敢拿出来说。还有两个组合，一个是基于规则处理浅层语言模型的，比如做英语的词法的，词根的变换，还原等，是可以做一点但有限，油水不多了。基于统计处理深层语言模型，是有的，但由于深层模型的本身是有一个资源建设这样一个积累在里面的。如果没有资源建设，深层是无从谈起的。你做统计，又要有语料的积累。语料和资源建设的双重积累是瓶颈，能够用打通瓶颈往前走的人不多，但我认为这是正确的路。



在自然语言处理领域，我感觉到一个不好的现象：厚此薄彼。我们看到，理工农医各类专业的领域专家是很受尊敬的，唯独语言学家不受尊重。不仅很多和自然语言相关的大型的项目里面语言学家的参与度很不够，就连谷歌的诺维格和语言学家乔姆斯基之间的争论也是很不友好的，语言学里面有很多宝贝，但是外面的人不认识，认识不到。那为什么自然语言处理还能这么火呢，我认为这里面一个原因是他们还在吃我所说的“规模红利”。当我的语料规模非常大、应用面非常广的时候，即使是浅层，也能做出很好的效果来，为一些处理到浅层就够了的应用提供了很好的支持，所以会有这样的现象。但是另一方面，深层的玩法并没有大的突破，要想抄底，别人不做，你来做，总要有一些瓷器活，但是这个瓷器活目前还不行。



对目前这个主流的基于统计的浅层自然语言处理，有两点我认为是需要高度肯定的：

第一是使用正确的手段解决了语言边界这样一个典型的非良定义的人工智能的问题。所谓非良定义，就是说人知道这个边界在哪，但是没有数学公式把这个边界写出来，而且这个边界是上下文相关的，在不同的环境下是浮动的。对这样的东西，使用统计学习的方式来获取领域知识，这是正面的，我认为这个方向是对的。

第二，是把符号的根基，也就是语言中不可再分的符号代表了什么这件事情，映射到，或者说植根于网络空间之上。说几个简单的例子：关于你在哪的问题，跟你手机里的地图联系起来了；关于你跟哪个人什么关系的问题，跟你的通讯录联系起来了，跟你的社交网络的账号联系起来了；天气的问题，甚至是车次的问题，跟相应的这样一些服务联系起来了。这个是一个最重要的、值得高度肯定的一点。



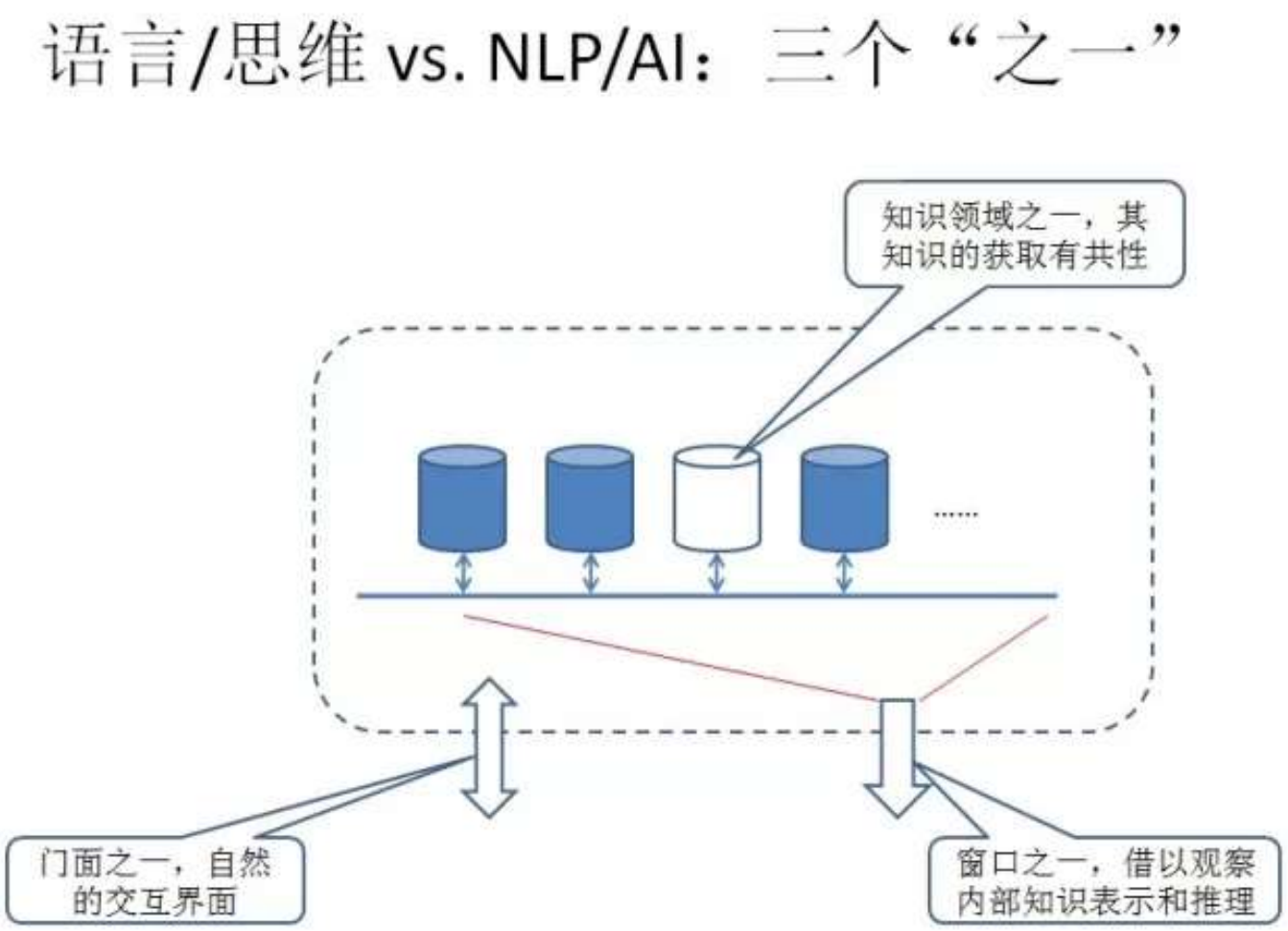
但是底下我们还是要讲，这是不够的。对语言学领域的这个知识到底长什么样，如果不知道长什么样就想学习，或者说随便攒一个长什么样，然后就学习，这个效果会差得很远，很少有人真的去关注语言学知识真往深了走到底长什么样。深层的知识躲在后面，似乎没有什么硬标准能够把它们拎出来，或者说你可以画一样，他可以画另一样，大家的画法不一样。但是，大家要解决、要共同面对的问题难点是明确的：一个是递归嵌套的深度，一个是语言成分远距离相关的宽度。相关成分相距越远、递归嵌套层数越深，浅层模型越难处理。只要这两个难点有所突破，我相信不管是把语言模型画成了一个什么样子，这个样子肯定在某种程度上反映了深层的语言学知识。



我个人在自然语言处理领域的研究兴趣，一是关注语言学知识长什么样，就是刚才所描述的问题；二是利用RNN（递归神经网络）做语言知识的自动获取，这个有很多的花样，比如可以加栈、加计数器，可以把规则编译成RNN等等，和自动机的机理非常接近，这个领域是值得特别关注的；三是用自然语言作为知识表示直接进行模式推理和检索，是我带着学生做的方向。个人在纯语言学方向也有一些研究兴趣。



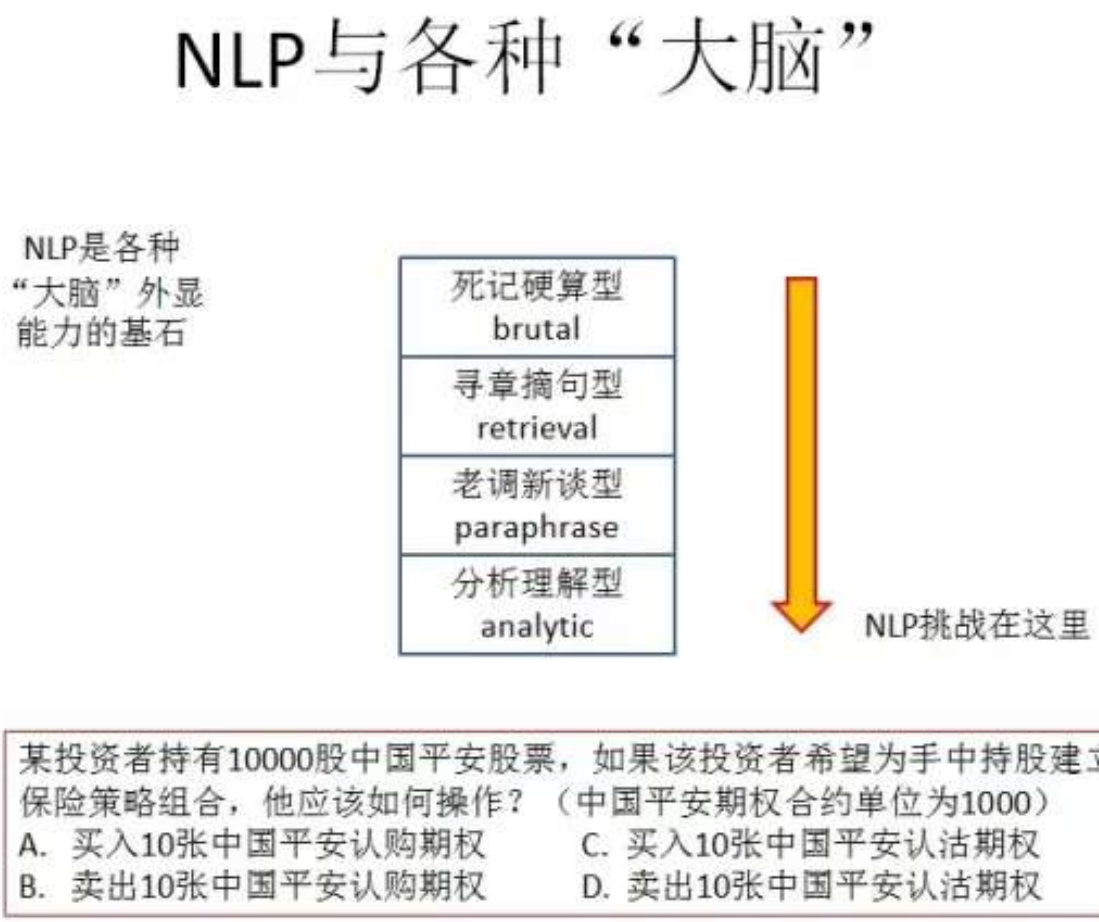
说到语言与人工智能的关系，我认为语言从三个角度对人工智能是有贡献的：第一，语言是一个自然的交互界面，善解人意绕不开这个界面；第二，语言背后是一套知识，对它怎么学习，怎么表示，和对一般的知识怎么学习和怎么表示，是有共性的；第三，语言是一个窗口，通过对人类语言的观察，我们可以侧面了解人类大脑内部发生的推理和表示。



但我还是要提两个“不指望”。第一是我们做自然语言处理的，不指望人工智能的天上掉下语言处理的馅饼，还是靠自己认识语言，真正地在这个语言上做文章，而不要在其他的地方做文章，其他的地方没有出路。第二是不指望语言处理的这个方向，如果突破了的话，会导致人工智能领域出现什么样的飞跃或者奇迹，这也是不现实的，因为语言这个东西就是刚才说的许多类别的知识领域之一，就是一个边角的东西，虽然很重要，虽然躲不开，但是对人工智能那边没有太大的影响，那边该怎么研究就怎么研究。

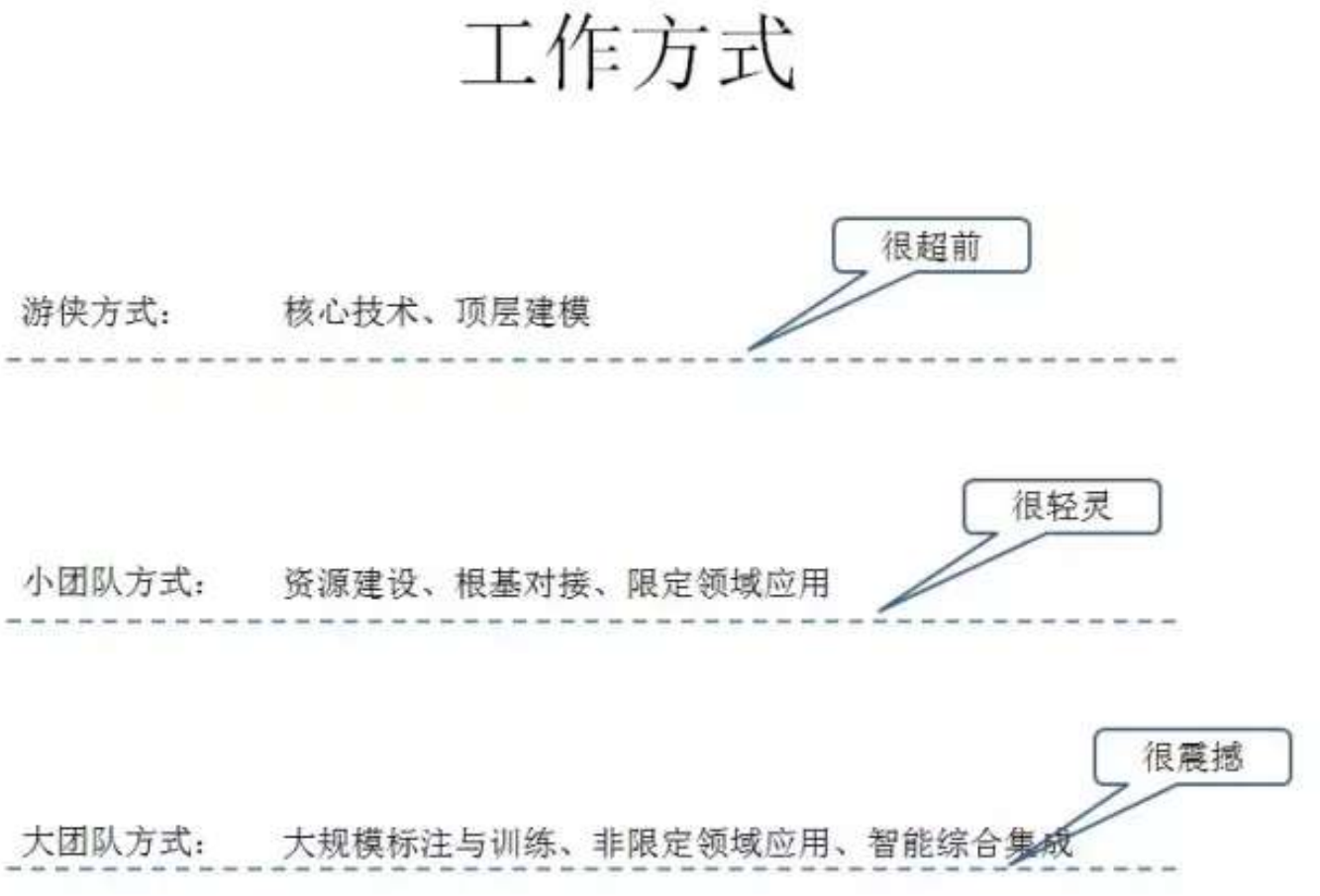


看到各种大脑计划，大脑计划的外显能力都离不开自然语言的处理，但是我们还是可以根据语言处理所用到的技术的不同，把它们分几个类型，如下图所示：



其中刚才说的高考是属于第三个层面的，最后一个类型（分析理解型）最难，其典型就是应用题。谁能解决了我觉得很令人佩服，如果真做好了说明是对语言深一层的东西掌握了。上图所举的应用题，是我们期权培训教材里面一个考题，自然语言理解怎么样处理这样一个题，通过什么方式入手来处理，很有讲究。我们看到这里面有计算和分析，还有价值取向：要赚钱。在这道题里面隐含了一个答题目标，就是如何做才是赚钱的。这是典型的投资逻辑。

搞大脑计划，一般都是一个非常庞大的团队，会进行大规模的标注和训练，对非限定领域的大数据大语料进行训练和开发，对各种智能化技术进行综合集成，这个做出来会很震撼，很酷。但小团队也有优势，可以比较轻灵，在特定领域，限定领域作出出色的应用，比如出门问问。还有另一种方式，就是游侠，他们不太可能大规模地推开业务，只会做一些核心技术，特别是顶层的建模。我一直主张，在一个自然语言处理系统里，语言学知识到底长什么样，决定了把后续的学习手段加上之后到底能够走多远，这个东西我想一个项目立项之时，这个思路就已经定了，后面再做也超不过先天局限。而恰恰是这个游侠方式，可以在这个地方走得更远。



总结一下我的主要观点：自然语言处理的核心关键问题还没有解决，但应用方面取得一些进展不是偶然的，有其走得对的地方。突破的钥匙掌握在语言学家或者是通晓语言学成果的人手里。最要害的试金石是两个：递归嵌套，远距相关，如能做到，说明对语言的深层处理能力有标志性的进步。我们认为后面有很多机会，不仅大脑计划有机会，游侠模式也有机会，统计走不下去的地方，规则抄底的机会也是存在的。最后说与人工智能的关联。虽然我是这样的题目，结论是自然语言处理和人工智能并没有强关联，而自然语言处理可以为人工智能的进步做一点点贡献，但是是比较微弱的。谢谢。

附：讲者简历

【白硕】博士，研究员，中国中文信息学会常务理事，信息检索与内容安全专业委员会主任委员，中国科学院计算所、信工所和中国科学院大学兼职博士生导师，上海证券通信有限责任公司董事长。曾长期从事自然语言处理、信息检索相关领域研究工作。目前主持轻便高效证券交易系统和证券行业云服务等证券领域信息技术应用研究。

« 上一篇 下一篇 »
分享到：

您可能感兴趣

人工智能本可避免德翼空难？
世界经济论坛最新科技榜单：2015年十大新兴科技
是人还是机器？智能测试和图灵测试完全不一样！
为什么说人工智能能实现弯道超车
人工智能技术将如何颠覆产品设计和提升用户体验

热门阅读

一天内 一周内 一月内

刘强东：我管75000人靠这4张表格
有钱人就一定会投资？未必！
中国土豪特供：林肯大陆新款闪亮登场
活着，不是一个人的事
像“王牌特工”一样穿西服

我要点评 **+ 更多评论（0）**

更多商业资讯，安装『福布斯中文网APP』



[10.205.160.37:4002] HTTP/1.1 400 Bad Request

1 北京到台北机票携程2折起
2 信用卡必学：提升技巧课程培训

Copyright@2010Forbes.com Inc福布斯公司版权所有
沪ICP备09033453号