## 机器学习：人工智能的实现形式

当前实现人工智能的主要方式即机器学习。顾名思义，机器从数据中学到种种规则，有些规则符合人类认知，而有些规则却超出人类认知：这好比一个现代人在古代象形文字中摸索规律，试图破译语言理解的密码，但它往往并未理解数据自身蕴藏的知识，规则结论均来自于数理统计等方法。机器学到的规则更客观、更全面、更具“突破性”，这一点在大数据挖掘上就有着很好的体现。但在模拟人类感官和大脑智能时，机器却经常自作聪明、画蛇添足，此时机器在数据集上的表现往往被称为过拟合，是开发者所不期待的。这一特性与近些年实际的发展完全相符：传统侧重特征工程的机器学习方法在数据挖掘上表现抢眼，所以AI时代首先出现了大数据、云计算的概念；而模拟视觉、听觉、语言理解等方面的智能直到CNN（卷积神经网络）、RNN（循环神经网络）、Attention（注意力机制）等深度学习方法崛起才逐一取得了突破进展，深度学习随后也迅速流行起来，现今已是人工智能最重要的实现方式。展望未来，毫无疑问深度学习的风向标会指向GNN（图神经网络），图结构具有最强大的表达能力和可解释性强的优势，因此对GNN的研究极有可能会为当前过于单调冗余的深度学习领域找到突破口；此外，鉴于以监督学习为主体的深度学习十分依赖带标签的数据和强大的计算能力，因此无监督学习、强化学习等领域的进展也非常值得期待。

从技术层面来看，机器学习到底是怎么实现的？简单地说，常规的机器学习就是对一个可以表示成流程图的任务进行最优化的过程。对监督学习来说，优化任务完成后会得到一个模型，随后可用该模型对未知问题进行预测得到结果，因此优化任务又被称为训练过程；对无监督学习来说，优化任务直接得到结果，所以通常无监督学习没有训练的过程。监督学习中得到的模型是指一种最优的明确的处理方案，一般表现为具有最优参数的复杂的表达式，这个式子的最终解便是我们想要的结果，或是某种结果的概率。有时模型一词也会出现在无监督学习中，一般用来表示优化完成后得到的向量或矩阵。换言之，常规的机器学习实际上就是数学的应用。

再来对比一下机器学习和编程。一个任务的决策过程往往可以一步步地展示出来，否则不能称之为逻辑决策，而是随意或凭空的猜测。上文已经提到监督学习是使用训练出来的模型进行决策，即选择一个类别；而编程则经常是通过若干循环体和if-else结构，在层层遍历与判断之后得出结果。编程中的各种条件来自于开发者的主观设计，堆叠出大量的条件选择也是一件非常耗时的工作，但是无漏洞程序的最终结果是可以期待的，因为它不会出现意外。与之对比，在很多机器学习算法的实现上，机器学数据的过程就类似于编写了上百万甚至更多的if-else条件判断，而被触发的条件所对应的处理为最终的类别选择起到了贡献（线性CRF的实质即是如此）。实际上，机器学习学到的规则大部分都是无法使用if-else显式地编写出来的，但即便能够编写往往也需要难以想象的代码量，很显然这体现了机器学习的智能性；但对于某些编程即可解决的任务，用机器学习就显得大材小用，且问题的输入稍加修改还可能会导致模型分类错误。所以，要适当地应用机器学习，并非所有决策任务都需要机器学习才能解决；在编写if-else就可以解决任务时，编程得到的结果一定比机器学习的分类结果更加可靠。

## NLP：人工智能的重要分支

NLP（自然语言处理）是人工智能的一个重要分支。它借鉴人脑的思维模式，致力于通过对文本数据的处理与学习，进而接近人类理解文本并获取知识的能力。随着人工智能时代的到来，国内外科研机构及科技公司高度重视NLP的发展，NLP技术如雨后春笋般层出不穷。现有的技术成果已可对文本挖掘等任务提供巨大帮助，而未来NLP的发展重点将落在深刻理解语义、知识推理等方向上。可以预计，在未来数十年，NLP都会拥有光明的前景，作为科技的最前沿引领AI时代的潮流。

NLP任务主要使用深度学习（深层神经网络）和概率图模型两大理论体系。而近来深度学习在机器翻译、文本分类等方向上的强势崛起更是有力地推动了NLP的发展。深度学习采用具有连续特性的分布式向量或矩阵来表示各种事物（图像、语音信号，文本信息，实体关系等等），而逐渐摆脱具有局限性的离散化特征。此外，深度学习还追求处理手段的统一化：无论是哪种模型，其本质都是在最优化的任务过程中，通过深层网络不断编解码（张量变换），最终达到高水平的拟合效果。从结果上看，深度学习可以从海量数据中更加精准细致地学到特点规则，克服了基于传统特征工程中任务质量严重依赖特征设计的缺点。但是，深度学习当前也存在可解释性差，对数据规模的要求更高，模型过于“庞大”、“臃肿”等问题。

概率图模型是另一种广泛应用在NLP上的理论体系。相比于深度学习，概率图模型具有更加严谨明确的理论支撑，可解释性良好；通常在规模不够大的数据集上完成相同任务时，基于概率图模型的算法效果要好于深度学习方法。概率图模型是概率统计上的精华，在NLP领域中最常用的两类概率图模型是HMM（隐马尔科夫模型）和CRF（条件随机场）。其中，后者可以解决多样的问题，且总能得到不错的效果。实际上，目前广泛使用的线性链CRF仅仅是该理论体系中最简化的方法，概率图模型的应用还可以进一步扩展。但是概率图模型严格从数学角度出发，对理论基础的要求非常高，且复杂概率图的实用性也有待探究，这些都阻碍了概率图模型体系的发展。

在NLP技术上，基于CRF的中文分词早在几年前就已达到商业化的水准，至今仍被视为最好的汉语分词方法。而基于深度学习的机器翻译在近几年不断突破，也具有了商业化水准。最新最有代表性的产品莫过于Google Brain 开发的基于Attention的模型Transformer。基于分布式表示的Word Embedding（词嵌入）是深度学习在NLP上开花结果的关键，这种借助词的语境及语料统计得到的词向量不仅可以用作神经网络的输入，还可用于词义分析。以Word2Vec、GloVe、FastText为代表的方法摆脱了深度学习对算力需求量巨大、模型沉重等固有弊病，用轻量级的神经网络便可快速得到质量有保证的词向量或分类模型。相比之下其它核心任务（如词性标注、句法依存分析、情感情绪分析）目前还未有显著性的突破。此外，汉语与英语NLP的发展也存在着明显的差距，这主要是由语言本身差异、汉语语料不足以及研发人员数量和水平上的落后引起的。我国的NLP发展任重而道远！

如果把上述提到的NLP技术看作是“上半场”，那么更加关键的“下半场”就是知识图谱。在AI中，如果说图像、语音识别对应着人类视觉、听觉功能，基础NLP任务对应着人类对外界知识讯息的简单处理，那么知识图谱则对应着人类对知识的总结理解能力以及基于已储备知识的逻辑推理能力。显然，知识图谱作为NLP的进阶任务，能够实现更高级的智能，真正的智能决策必定离不开高质量的知识图谱。与其重要性相应，目前GNN已经成为深度学习领域的研究热点。所以，不夸张地讲，GNN是深度学习的未来，知识图谱是NLP乃至人工智能的未来！

## 语料库与词库：NLP的依赖材料

既然机器学习的本质是学习数据，那么在训练任务时就必须要有数据可学。数据的类别在不同领域任务上各不相同，在CV（计算机视觉）方向图像便是数据，而在NLP方向，训练数据统称为文本语料。文本语料还可以再细分成生语料和熟语料，生语料即最直接的文本，而熟语料，顾名思义，是经处理后添加了特定任务类别标签的文本。一般来说，计算机视觉等领域的类别标签更容易标注，甚至还可以通过互联网雇佣世界各地的人来打标签。但是对NLP而言，因为语言的差异，能够胜任某种语言文本标注的人群基数必然远不及图像等标注，而且在大多数NLP任务上还需要标注者拥有良好的语法基础且在标注前接受过一定的培训，所以构建NLP领域的语料是一项庞大耗时的工程。

由于历史因素等原因，汉语NLP的成熟度远不及英文NLP。目前，在各项NLP任务上，英文带标签的语料都比较充足，相比之下每种任务的汉语语料都捉襟见肘。其实这不仅仅是起步较晚造成的差距，语言之间的差异性也是原因之一。英语语法结构更加丰富具体，词与词的分隔非常明确，这保证了英文在非语义的层面上歧义性不大。而汉语文本中词与词的分隔必须通过人来进行决断，且汉语的歧义性更强。当然，汉语在NLP上也有积极的一面，汉语的常用字词数远少于英语，句法表达上也远没有英语这么复杂。这些特点与深度学习的设计理念很相符，因此未来高水平的深度学习模型很有可能在汉语言处理的任务上达到非常高的水准。

要想大规模构建高质量的汉语熟语料，词分割的规范是必先要解决的问题。语料的规范与其本身内容之间是牵一发而动全身的关系。规范一旦被修正，通常只有通过人工修改内容才能保证语料的质量，但无止境的重复劳动无疑是对人力资源的浪费。在NLP中，一切任务都是在词语被分割后进一步实现的。但遗憾的是，在汉语词分割上并没有一套权威统一的标准。一些非盈利机构虽对其部分语料进行了开源，但语料规模有限，语料来源分布也较单调，很多语料因年代久远而无法很好地诠释当今主流词法，过去的某些规范因技术的革新可能也已不再合适。总之，权威词分割规范的欠缺是汉语语料构建最致命的问题！

当然，词分割规范不统一并不代表没有规范，事实上各份开源语料的分割规范均不一致。但只要能统一当前使用的这份文本语料的规范，便可用以训练模型并评估方法的效果。因为机器学的是任意数据的规则，所以这些开源语料均可以用在学术研究类的工作上。然而，开源语料由于规范不合理等事实在商业化使用上出现了问题，其根源在于商业使用时效果的“评估者”由单纯对比所有标签和全部预测结果得到的偏差变成了用户的使用感受。

词库用来收录词语及其部分重要特征（如词性、词频），是逐渐积累的产物。与语料相比，词库并非必不可少，但高质量的词库会在某些任务中起到辅助性的作用。举例来说，在分词任务中，挂上一个长词（如成语、专有名词）词库会大大提升分割长词的准确性；在词性标注任务中，在处理的最后环节搜索一遍词库也会使词性预测准确率提升，特别是针对具有单一词性的词汇。现有的用于NLP基础任务的方法是很成熟的，在具有突破性的方法浮现之前，开发机构比拼的就是语料、词库的规模和质量。实际上，已完成分词的语料也可以看作是词库，只须对这份语料稍加处理便可得到一份不重复的词集。

## NLP在金融领域的应用价值

毫无疑问，人工智能是引领未来的战略性技术，世界各国都在加速构建本国的AI生态。2018年上海人工智能大会举办了以“智能金融：定义金融新时代”为主题的论坛，可见未来的金融发展定会与人工智能技术紧密联系。而NLP就是金融领域在引入人工智能概念时最直接、最有意义的工具。

NLP中的文本纠错技术可以帮助人们快速发现各类金融文件、文本信息在语言运用上的错误（如错别字、词语的谐音错误、句法语病等）。情感类分析任务中，可以通过机器快速得到文章或关键语句的情感甚至情绪，而这类信息极有可能会对接下来的金融策略起到关键性的作用。人们在做决策时往往会回顾历史，在大量相似历史事件中获取经验；这种经验在计算机上可以表现为一种事件与关系的数据结构，而通过NLP的句法分析可提取句子主干成分，再进行一些必要的处理，便可以构建这种历史事件关系库。知识图谱更是NLP的“终极武器”。构建领域知识图谱后，甚至可以实现诸如智能选股、因果分析这样的工作。

目前，虽然NLP无法提供准确可靠的分析结果，但退一步讲，用计算机进行文本处理本身就已经给个人的工作带来极大的帮助。个人精力有限，而机器可以不知疲倦地工作。因此，即便忽略高端的智能，而只是把NLP技术应用于工作中的一些基础任务上，同样也可以提高个人的工作效率。如今获得海量数据信息的代价越来越低，把NLP工具当作一种辅助，借助机器计算能力强的特点，对大数据进行统计分析，再辅以人的逻辑和经验进行人工决策，可以同时充分地发挥人和机器的优点，是当下运用机器智能最理智的选择！

在一个没有错误的程序下，机器可以永不出错地完成某一特定的任务，这是传统的IT。而AI时代则不同：机器学习旨在学习数据的规则，然而任何一份训练语料都不可能提供该种数据的全部规则，所以当遇到机器未曾探寻过的模式时，机器必然会犯错。当前AI最显著的特点就是常见模式不犯错，一旦犯错，不是错在重要问题上，就是在敏感问题上得到了让人哭笑不得的结果。因此，目前还鲜有金融机构敢于完全使用AI模型进行重大决策，即便他们已经开发出了很先进的模型。惨痛的犯错代价可能会导致AI模型无法得到像在图像识别或机器翻译等领域一样的广泛应用，或许股民还是更愿意相信自己的选择而不是智能系统推荐的股票。可以预计，AI在金融领域的发展必然会举步维艰，但这并不意味着AI在金融方向无能为力。展望未来，或许真正能够在金融领域发力的不是当下最热门的深度学习，而是强化学习或自适应学习。换一种说法，毕竟深度学习只是实现人工智能的一种途径，我们不能期待深度学习这把看似万能的钥匙一定能打开金融领域牢固的枷锁，或许我们要回溯到人工智能的初衷，再去探寻突破口。

综上所述，NLP在金融领域有着不错的应用价值，更有着巨大的开发潜力！NLP是最能与金融紧密结合的人工智能，知识图谱工程尤其显示出了这一点。历史事件总是高度相似，常用的推理思想和过程毕竟有限，设计合理的算法进而用程序实现并非是不可能完成的任务。因此，我们有理由相信以NLP为代表的人工智能技术在金融应用上不会止步于华而不实的理念；在未来，质量可靠、具有一定规模的金融图谱必将在金融分析上大放异彩！