Kenneth Church：钟摆摆得太远——人工智能和自然语言的发展与未来

> 译者按：肯尼斯·丘吉(Kenneth Church) 是自然语言领域的泰斗，语料库语言学和机器学习的开拓者之一。丘吉的这篇长文《钟摆摆得太远》(A Pendulum Swung Too Far) 是一篇主流反思的扛鼎之作。作者在文章中回顾了人工智能发展中，理性主义和经验主义各领风骚此消彼长的历史规律，并预测了今后20 年自然语言领域的发展趋势。文章的主旨是，我们这一代学者赶上了经验主义的黄金时代（1990 年迄今），把唾手可得的低枝果实采用统计学方法采摘下来，留给下一代的都是“难啃的硬骨头”。20 多年来，向统计学一边倒的趋势使得我们的教育失之偏颇。现在应该思考如何矫正，使下一代学者做好创新的准备，结合理性主义，把研究推向深入。丘吉的忧思溢于言表。丘吉预测，深度网络的热潮为主流经验主义添了一把火，将会继续主导自然语言领域十多年，从而延宕理性主义回归的日程表。但是他认为理性主义复兴的历史步伐不会改变。他对主流漠视理性主义的现状颇为忧虑，担心下一代学者会淹没在一波又一波的经验主义热潮中。

实用主义动机

20 世纪90 年代，经验主义的复兴是一个激动人心的时刻。我们从来没有想到，我们的努力会如此成功。当时，我们想要的只不过是一席之地而已。在当时流行的各项研究之外，我们所想的只是为不同于当时其他研究的工作争取一点空间。我们成立了SIGDAT为这类工作提供一个论坛。在1993 年成立之初，SIGDAT只是一个相对较小的关于大语料库的会议论坛，后来演变成规模较大的EMNLP 会议。起初，SIGDAT 会议在很多方面（规模、主题和地理范围）都与主流ACL大会非常不同。然而若干年后，这些区别已经很大程度上消失了。两个会议靠拢，这让人感到高兴。但我们可能是太成功了，我们不仅成功地让我们感兴趣的工作登堂入室，没给其他工作留下多少空间。图1 展示了从理性主义到经验主义的这一戏剧性转变。这种转变还在继续，似乎看不到尽头。

根据霍尔(Hall) 等人的文章，这种转变始于1988 年布朗 (Brown)和丘吉的工作。霍尔等人的依据是对ACL 文集的分析，文献包括自20 世纪70年代至今在计算语言学领域发表的总计16500 篇论文。

但是，如果我们考虑一个更长的时间段，追溯ACL 文集以前的文献，我们看到的是一幅非常不同的画面，如图2 所示。更加显著的趋势是经验主义与理性主义之间的振荡，像钟摆一样，每隔二十多年来回振荡一次：

- 20世纪50 年代：经验主义（香农(Shannon)、斯金纳(Skinner)、弗斯(Firth)、哈里斯(Harris)）；
- 20世纪70 年代：理性主义（乔姆斯基(Chomsky)、明斯基(Minsky)）；
- 20世纪90 年代：经验主义（IBM 语音团队(IBM Speech Group)、AT & T 贝尔实验室(AT&T Bell Labs)）；
- 2010年代：回归到理性主义了吗？

本文将回顾一些我们这一代人曾经“反叛”的理性主义观点。遗憾的是，我们这一代是如此成功，以至于这些理性主义观点被人们忘却了（如果我们接受图2给出的预测，那么现在正是理性主义应该复苏的时

期）。有些重要的理性主义代表人物如皮尔斯(Pierce) 在当今流行的教科书里甚至没有提及。如此下去，下一代人可能没有机会听到理性主义一方辩论的声音。特别是，如果理性主义立场在今后几十年逐渐流行，理性主义者可以提供很多值得重视的见解。

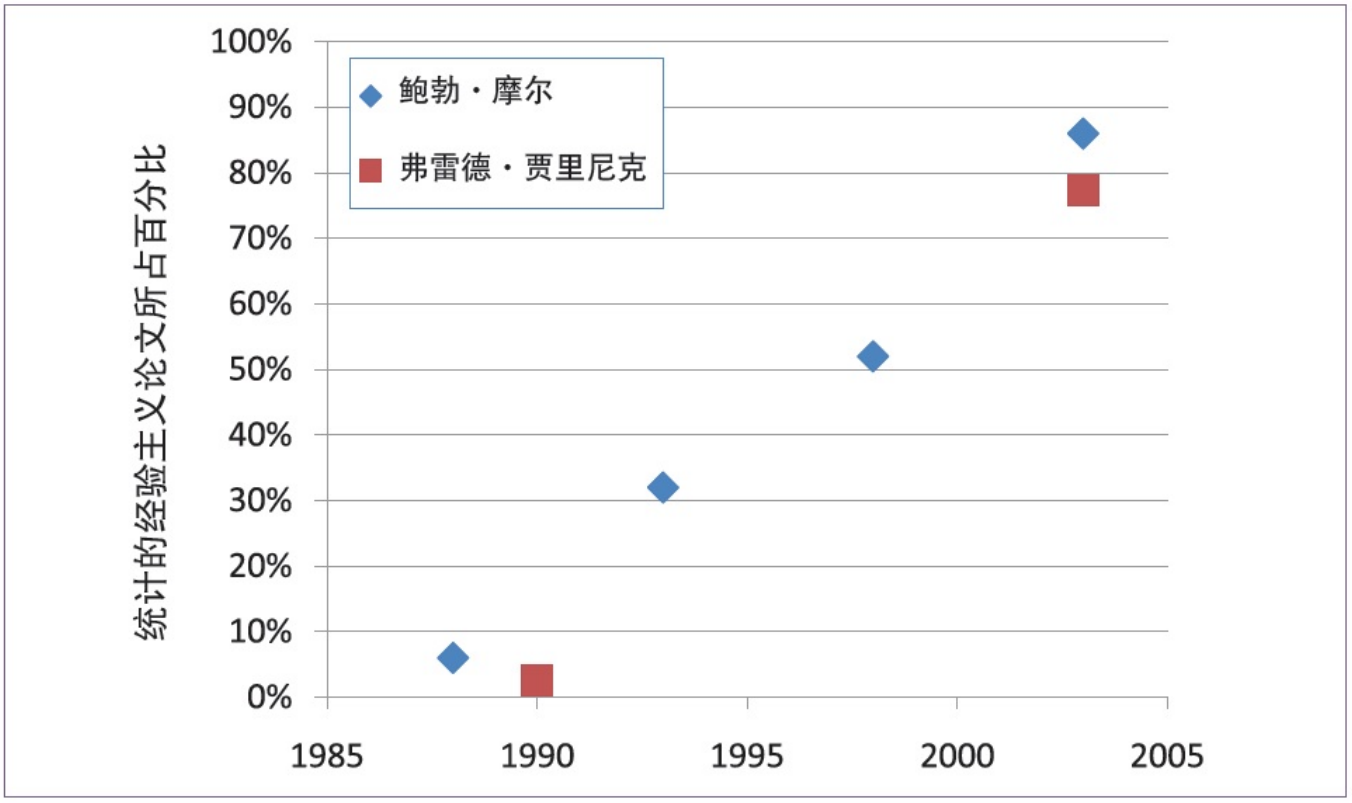


图1 理性主义到经验主义的转变令人惊讶（而且毫无争议）。该图是基于鲍勃·摩尔(Bob Moore)和弗雷德·贾里尼克(Fred Jelinek)对ACL会议的独立调查（私人通信）

是什么促使20 世纪90 年代经验主义的复兴？我们当时在反抗什么？经验主义复兴实际上是受到了实用主义考量的推动。学术界当时正埋头研究自然语言中面临的巨大挑战，例如完备人工智能(AI-complete)的难题和远距离的依存关系。而我们所提倡的是从务实的角度来先针对一些较简单的、较有可能求解的任务，例如词性标注。当时数据的获得变得前所未有的方便。我们能用这些语料数据做些什么呢？我们认为，做成一些简单的事情比根本不做强。让我们去摘取一些低枝的果实，让我们利用近距离依存关系做我们能做的事情。虽然那不能解决整个问题，但还是让我们专注于我们能做什么，而不是我们不能做什么。玻璃杯有一半是满的（而不是已经空了一半）。

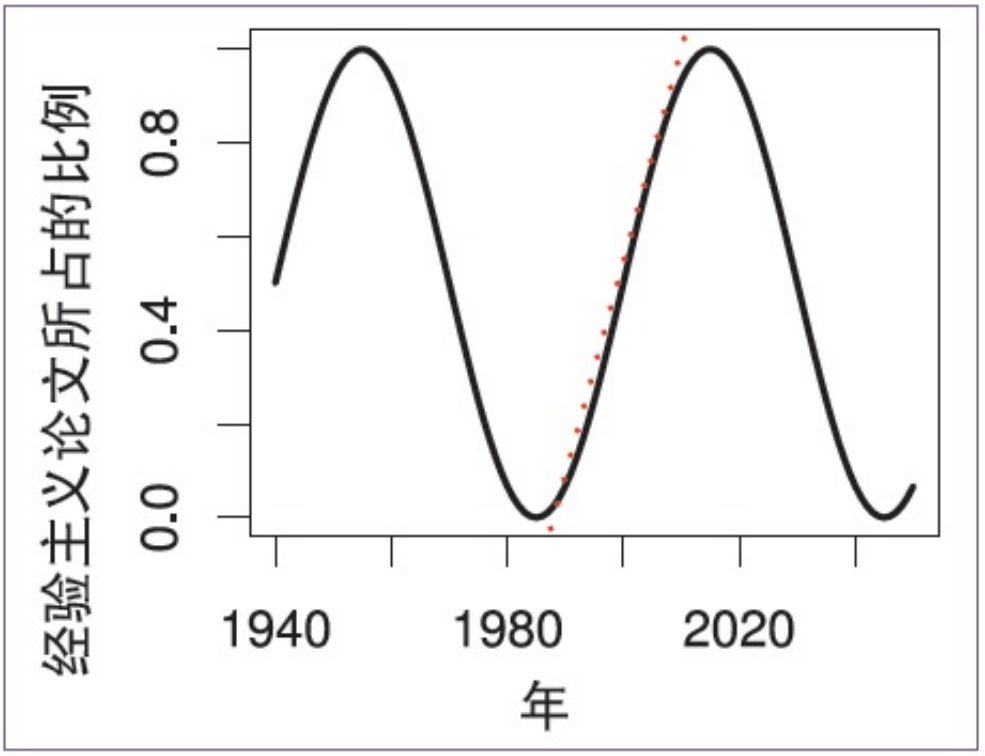


图2：对文献不寻常的解读，其中图1所示的趋势（此处以红点表示）是每隔20多年更大振荡的一部分。注意红点所示的是实际数据，而振荡曲线所示意的趋势只是为了说明一个观点

我们当时是这样记述这段历史的：

> “20 世纪90 年代重现了具有20 世纪50 年代风格的语言分析的经验主义及其统计方法。50 年代是经验主义的高峰期，主导了从心理学（行为主义）到电子工程（信息论）一系列广泛的领域。当时语言学的通行做法是，不仅仅依据词义，还要基于它与其他词共同出现的情形来划分词类。50 年代英国语言学领域的领袖人物费思(Firth)用一段令人难忘的话总结此方法：‘通过一个词周围的词来了解这个词的意义。’遗憾的是，受一系列重大事件的影响，50年代后期和60年代早期，经验主义式微。这些重大事件包括乔姆斯基(Chomsky) 在《句法结构》(Syntactic Structures) 中对N 元文法 (n-grams) 的批判，明斯基与帕佩特(Papert) 对神经网络的批判。
>
> 经验主义复兴最直接的原因也许是大量数据可用：文本从来没有这么丰富过。10 年前，搜集了100 万词的布朗(Brown) 语料库就被弗朗西斯(Francis) 和库塞拉(Kucera) 认为是大数据，但即使在那时，也有更大的语料库，如伯明翰(Birmingham) 语料库。如今，许多地方的文本样本已经达到上亿甚至几十亿词量……。通常称为文本分析的数据密集型语言研究方法采取的是实用主义手段，非常适合近来被强调的数值评估和具体的任务。文本分析强调对非受限文本(unrestricted text) 的广泛覆盖（尽管可能肤浅），而不是对于（人为）限定领域的深度分析。”

寒冬

20世纪90年代早期， 研究界发现应该注重务实方法，原因之一是该领域当时正处于严重的资金寒冬，史称第二季人工智能寒冬(AI winter of1987~1993)。在又一次资金萧条到来之际，研究共同体比较容易接受一种更加现实的、结果更可靠的新方法。根据维基百科资料：

> “在人工智能的发展历史中，所谓人工智能寒冬是指社会对人工智能研究的资助和兴趣消减的时期。许多新兴技术都经历了从狂热、失望到资金削减的过程（例如历史上的铁路大开发以及网络泡沫），

> 但是人工智能的问题更加突出。这种模式已经发生过许多次了：
>
> ● 1966 年：机器翻译的失败；
> ● 1970 年：放弃人工智能联接主义(connectionism) ；
> ● 1971~1975 年：美国国防部高级研究计划局(DARPA) 对卡耐基梅隆大学语音理解研究项目的失望；
> ● 1973 年：莱特希尔(Lighthill)人工智能评估报告(Lighthill Report)发表之后，英国对人工智能研究资助的大幅削减；
> ● 1973~1974 年：DARPA 削减对人工智能学术研究的资助；
> ● 1987 年：Lisp 机市场崩溃；
> ● 1988 年：战略计算规划(the Strategic Computing Initiative) 取消了进一步资助人工智能的计划；
> ● 1993 年：专家系统慢慢跌入低谷；
> ● 1990 年代：第五代计算机项目的原始目标黯然淡出视野，以及被牵累迄今的人工智能的坏名声。
>
> 人工智能经历的最糟糕的时间段是1974~1980 年和1987~1993 年。有时人工智能寒冬指的就是两者之一（或两者的某个时间段）。”

寒冬常常紧跟着过度的乐观主义，例如西蒙(Simon)在文献[30]中提到的：

> “在不久的未来——不会超过25年——我们将会有技术能力用机器来代替机构中的任何人类功能。而且，我们将充分掌握人类认知过程及其与人类情感、态度和价值观的交互过程的理论，这些理论将会被实验所证实。”

如今，比起第二季人工智能寒冬，我们变得更有信心。15 年低枝果实的采摘已经取得了相对稳定的成果，也获得了相对稳定的资助，至少比人工智能寒冬的形势乐观很多。

皮尔斯、乔姆斯基和明斯基

毋庸讳言，我们所反抗过的伟大的理性主义者如皮尔斯、乔姆斯基和明斯基(Pierce, Chomsky and Minsky, 以下简称PCM)，对人工智能领域的现状不会感到满意。当然，另一方面，今天此领域的领军人物大多也不乐意看到PCM 理性主义的复兴。一位领域的带头人听说我在写这篇文章，讥讽道：“皮尔斯对我们现在有什么意义？”PCM 的观点在当年就饱受争议，现在依然如此，因为它们导致一些领域包括语音、机器翻译和机器学习多次进入了严重的资金寒冬。

本文主要感兴趣的是PCM三位大师理性主义的共同主线。不过也必须指出，这三位大师的声音并不完全一致。在信息论方面他们有很大分歧。皮尔斯对香农和乔姆斯基二位均大加赞佩，尽管乔姆斯基对香农在信息论方面的许多工作持反对意见。很显然，这些观点并不能清楚地划分成不同学派（例如理性主义和经验主义），学派之内并非完全一致，学派之间也不是处处相异。

关于智能亦有很多不同意见。明斯基是人工智能的创始人之一，而皮尔斯一直是直言不讳的批评者之一。他说：所谓人工智能真乃愚蠢之极7。皮尔斯反对任何试图接近人类智能的东西，当然包括人工智能，也包括机器翻译和语音识别。皮尔斯主持了著名（或者说是臭名昭著）的语言自动处理咨询委员会(Automatic Language ProcessingAdvisory Committee, ALPAC) 报告。这一报告直接导致了机器翻译的资金寒冬[27]。皮尔斯也曾为《美国声学学会会刊》(JASA) 撰写富有争议的通讯“语音识别往哪里

去”(Whither Speech Recognition?)，给语音识别研究的资金造成令人寒心的困境。

本文重在回顾他们的共同主线，而不是他们的分歧。PCM 对当年流行现今复兴的一系列经验主义方法，均提出过挑战。他们的反对意见对于许多当今流行的方法都有影响，包括模式匹配、机器学习（线性分离机）、信息检索（向量空间模型）、语言模型（N 元文法模型）和语音识别（隐式马尔可夫模型(hidden Markov models, HMMs) 以及条件随机场(conditional random fields, CRFs)）。

学生们需要学会如何有效地使用流行的近似方法。大多数近似方法基于简化的假设，这些假设在多数情况下有用，但并非万能。例如，N 元文法能捕捉许多依存关系，但当依存范围超过n个词距离的时候，N 元文法则无能为力。同理，线性分离机在很多情况下可以区分正例和反例，但对无法线性区分的样例自然无效。许多这类限制显而易见（由其本性所决定），但即便如此，相关的优劣争论有时仍然很激烈。有时候，争论的某一方不再被写进教科书，逐渐被遗忘，只能期待下一代学者去重新发现或复兴。

乔姆斯基论述了N 元文法的局限，明斯基论证了线性分离机的局限。也有学者对于其他近似方法的种种局限提出看法。例如，图基(Tukey) 教导学生如何有效使用回归算法[34]。他鼓励学生测试各种正态假设的偏离现象。离群点(outliers) 是回归算法常见的麻烦来源，正如偏离直线的弯曲残差(bowed residuals)。很多人提出了种种绕行的补救方案。一个常见的手段是对数据做非线性变换，如对数变换。这些技巧把问题转化为另一个问题，使其偏离假定的麻烦有所减少。

乔姆斯基的反对意见

如前所述， 乔姆斯基指出N 元文法不能捕捉远距离依存关系。虽然现在回想起来似乎是显然易见的，然而在当时，香农-麦克米兰- 布雷曼熵定理(Shan-non-McMillan-Breiman theorem)令人非常兴奋，对这条定理的解释是：在极限条件下，只要稍加制约，N 元文法模型足以捕捉字符串的所有信息（譬如英语的句子）。乔姆斯基认为，在极限条件下这条定理也许是正确的，但是N 元文法模型远远不是能概括许多语言事实最简洁的模型。在实际系统中，我们往往必须将N 元文法严格限制在某个（小的）固定的值k上（例如三元或许五元）。这种限长的N 元文法模型系统可以捕捉到很多语法关系一致性方面的现象，但并非全部。

我们应该将这场论辩教给下一代学者，因为他们可能将不得不比我们更加认真地对待乔姆斯基的反对意见。我们这代人很幸运，可以摘取到大量的低枝果实（也就是那些可以用较短N 元文法捕捉到的语言现象）。但是，下一代学者将没有这么幸运，因为在他们退休之前，那些捕捉得到的语言事实大多将被捕捉已尽，因此，他们很可能将不得不面对那些简单N 元近似方法无法处理的语言现象。

中心嵌套论(Center-Embedding)

乔姆斯基不仅反对N 元文法模型，也反对有限状态(finite state) 方法，其中包括很多目前流行的方法，如隐式马尔可夫模型和条件随机场。

有限状态方法超越了N 元文法，它不仅可以捕捉一切N元文法可以捕捉到的语言现象，而且可以捕捉超越N 词距离的语法依存关系。例如，下列文法表现了主谓在数上一致的关系，名词和动词应该一致，二者同为单数(sg) 或者同为复数(pl)。这样的文法可以捕捉超过N 词距离的依存关系。

S → Ssg
S →Spl
Ssg → NPsg VPsg

Spl → NPpl VPpl
NPsg → … Nsg …
NPpl → … Npl …
VPsg → … Vsg …
VPpl → … Vpl …

最大的问题是，此文法是否需要无限的内存。为了使这场辩论严谨，乔姆斯基引入中心嵌套的概念，并创建了现今被称作乔姆斯基层级体系(Chomsky hierarchy)的理论。

乔姆斯基层级体系不仅在语言学，在其他许多领域，例如计算机科学9，也具有非凡的影响力。克努特(Knuth) 坦承他在1961 年的蜜月期间读到乔姆斯基的文章，发现它是如此“奇妙的事情：在这个语言的数学理论中，我可以感受到一个计算机程序员的直觉”。

乔姆斯基指出，乔姆斯基层级体系与相应的生成能力之间具有一种简单的对应关系：

类型0 > 类型1 > 类型2 >类型3

递归可枚举文法 > 上下文有关文法 > 上下文无关文法 > 正则文法

特别是上下文无关文法可以涵盖并超越正则文法；有一些需要无限内存（栈）所做的事情，有限内存就做不到。乔姆斯基的论证是，中心嵌套是上下文无关与有限状态之间的关键区别。也就是说，当（且仅当）一个文法具备中心嵌套能力，它才需要无限内存（栈）。否则就可以用有限内存（有限状态机）处理。

表1 乔姆斯基层级体系

类型	文法	自动机
3	正则文法	有限状态机 (finite-state machines, FSMs)
2	上下文无关文法(context-free, CF)	下推自动机 (push-down automata, PDA)
1	上下文有关文法 (context-sensitive, CS)	线性有界自动机 (linear bounded automata, LBA)
0	递归可枚举文法	图灵机(Turing machines, TMs)

更正式地讲，如果一个文法中具有一个可以生成形如xAy的非终结节点A，其中x 和 y 均为非空，那么这个文法就是中心嵌套。如果x 或 y 为空，则可以得到较简单的左权或右权的嵌套。左权嵌套和右权嵌套均可在有限内存（有限状态机）中处理，而不像中心嵌套那样需要无限内存（栈）。

中心嵌套的一个简单例子是一个括号嵌套的文法：

< expr > → (< expr >)

括号嵌套文法是中心嵌套的一个特殊案例，其中 x 是左括号，y 是右括号。一个栈结构可以很容易地记录左括号与右括号之间的远距离依存关系，但这需要无限的内存。最大的问题是有限内存是否可以处理括号嵌套文法。乔姆斯基证明这是不可能的。更一般的表述是，有限状态方法无法捕捉中心嵌套。

乔姆斯基用下列样例论证英语为中心嵌套语言，因此超越了有限状态方法（如隐式马尔可夫模型）的捕捉能力。乔姆斯基假定英语具有一个非终结节点 S（代表句子或从句），其自生成的时候在它的左右两侧可以添加非空内容，如下所示：

1. S → If S, then S.
2. S → Either S, or S.
3. S → The man who said that S, is arriving today.

关于中心嵌套的语言事实一直存在争论。本文审阅者之一反驳中心嵌套所用的几点论证，我当年在我的硕士论文中也提过类似的质疑。语料库中很难找到超过两层或三层的中心嵌套11。不过，乔姆斯基的说法并非没有道理。想要描述上述语言事实，采用允许任意中心嵌套的文法较之采用仅有一两层中心嵌套的文法可能更容易和简洁。

到目前为止，N 元文法和有限状态方法等近似模型足够我们使用。虽然这些近似模型都有其明显局限性，但迄今难以找到更有效的替代方法。尝试捕捉不常见的远距离关系也许可以处理一些不常见的边缘案例，但它们带来的问题往往比解决的问题要多。工程师们发现，处理好常见的短距离依存关系比处理不太常见的远距离依存关系更为重要。至少，这是我们这一代人的体验。

尽管如此，我们还是应该为下一代学者做好准备，使他们有可能比我们做得更好。我们应该教给下一代认识目前比较流行的各种方法的长处和短处。他们需要了解我们所知道的最成功的近似方法，但他们也需要了解其局限性。下一代学者很可能会找到改进N 元文法的办法，甚至可能发现超越有限状态的方法。

明斯基的反对意见

明斯基和帕佩特表明，感知机（更广泛地说是线性分离机）无法学会那些不可线性分离的功能，如异或(XOR) 和连通性(connectedness)。在二维空间里，如果一条直线可以将标记为正和负的点分离开，则该散点图即线性可分。推广到n 维空间，当有n -1 维超平面能将标记为正和负的点分离开时，这些点便是线性可分的。

判别类任务

对感知机的反对涉及许多流行的机器学习方法，包括线性回归(linear regression)、logistic 回归(logistic regression)、支持向量机(SVMs) 和朴素贝叶斯(Naive Bayes)。这种反对意见对信息检索的流行技术，例如向量空间模型 (vector space model) 和概率检索(probabilistic retrieval) 以及用于模式匹配任务的其他类似方法也都适用，这些任务包括：

1. 词义消歧(WSD)：区分作为“河流”的bank 与作为“银行”的bank。
2. 作者鉴定：区分《联邦党人文集》哪些是汉密尔顿(Hamilton)写的，哪些是麦迪逊(Madison)写的。
3. 信息检索(IR) ：区分与查询词相关和不相关的文档。
4. 情感分析：区分评论是正面的还是负面的。

机器学习方法，比如朴素贝叶斯，经常被用来解决这些问题。例如，莫斯特勒(Mosteller) 和华莱士(Wallace) 的鉴定工作始于《联邦党人文集》，共计85篇文章，其作者是麦迪逊、汉密尔顿和杰伊(Jay)。其中多数文章的作者是明确的，但有十几篇仍具争议。于是可以把多数文章作为训练集建立一个模型，用来对有争议的文件做判别。在训练时，莫斯特勒和华莱士估算词汇表中的每个词的似然比：Pr(word|Madison)/Pr(word|Hamilton)。对有争议的文章通过文中每个词的似然比的乘积打分。其他任务也使用几乎相同的数学公式，如表2 所示。近来，诸如logistic 回归等判别式学习方法正逐步取代如朴素贝叶斯等生成式学习方法。但对感知机的反对意见同样适用于这两类学习方法的多种变体。

表2 朴素贝叶斯方法的4项应用

词义消歧	$score(context) = \prod_{Word\ in\ context} \frac{Pr(word\,|sense_1)}{Pr(word\,|sense_2)}$
作者鉴定	$score(doc) = \prod_{Word\ in\ doc} \frac{Pr(word\,|author_1)}{Pr(word\,|author_2)}$
信息检索	$score(doc) = \prod_{Word\ in\ doc} \frac{Pr(word\,|relevant)}{Pr(word\,|irrelevant)}$
情感分析	$score(doc) = \prod_{Word\ in\ doc} \frac{Pr(word\,|positive\ review)}{Pr(word\,|negative\ review)}$

停用词表、词权重和学习排名

虽然表2 中4 个任务的数学公式类似，但在停用词表(stoplist)上仍有重要的区别。信息检索最感兴趣的是实词，因此，常见的做法是使用一个停用词表去忽略功能词，如“the”。与此相对照，作者鉴定则把实词置于停用词表中，因为此任务更感兴趣是风格而不是内容。

文献中有很多关于词权重的讨论。词权重可以看作是停用词表的延伸。现今的网络搜索引擎普遍使用现代的机器学习方法去学习最优权重。学习网页排名的算法可以利用许多特征。除了利用文档特征对作者写什么进行建模外，还可以利用基于用户浏览记录的特征，来对用户在读什么建模。用户浏览记录（尤其是点击记录）往往比分析文档本身信息量更大，因为网络中读者比作者多得多。搜索引擎可以通过帮助用户发现群体智能来提升价值。用户想知道哪些网页很热门（其他和你类似的用户在点击什么）。学习排名是一种实用的方法，采用了相对简单的机器学习和模式匹配技术来巧妙地应对可能需要完备人工智能理解(AIcomplete understanding) 的问题。

最近有博客这样讨论网页排名的机器学习：

> “与其试图让计算机理解内容并判别文档是否有用，我们不如观察阅读文档的人，来看他们是否觉得文章有用。
>
> 人类在阅读网页，并找出哪些文章对自己有用这方面是很擅长的。计算机在这方面则不行。但是，人们没有时间去汇总他们觉得有用的所有网页，并与亿万人分享。而这对计算机来说轻而易举。我们应该让计算机和人各自发挥特长。人们在网络上搜寻智慧，而计算机把这些智慧突显出来。”

为什么当前技术忽略谓词

信息检索和情感分析的权重系统趋向于专注刚性指示词(rigid designators)14（例如名词），而忽略谓词（动词、形容词和副词）、强调词（例如“非常”）和贬义词15（例如“米老鼠(Mickey mouse)”16和“ 破烂儿(rinky dink)”）。其原因可能与明斯基和帕佩特对感知机的反对有关。多年前，我们有机会接触MIMS 数据集，这是由AT & T 话务员收集的评论（建议与意见）文本。其中一些评论被标注者标记为正面、负面或中性。刚性指示词（通常是名词）往往与上述某一类标记（正面、负面或中性）紧密关联，但也有一些贬义词标记不是正面就是负面，很少中性。

贬义词怎么会标记为正面的呢？原来，当贬义词与竞争对手相关联的时候，标注者就把文档标为对我方“正面”；当贬义词与我方关联的时候，就标注为对我方“负面”。换句话说，这是一种异或依存关系（贬义词XOR 我方），超出了线性分离机的能力。

情感分析和信息检索目前的做法不考虑修饰成分（谓词与论元的关系，强调词和贬义词），因为除非你知道它们在修饰什么，否则很难理解修饰成分的意义。忽视贬义词和强调词似乎是个遗憾，尤其对情感分析，因为贬义词显然表达了强烈的意见。但对于一个特征，如果你不知道其正负，即使强度再大也没什么用。

当最终对谓词- 论元关系建模时，由于上述异或问题，我们需要重新审视对线性可分的假设。

皮尔斯的反对意见

比起明斯基和乔姆斯基，皮尔斯在时下的教科书上更少提到，尽管皮尔斯作为ALPAC 委员会主席以及著名的“语音识别向何处去”一文的作者对本领域有深远影响。无论从终结资助还是从文章的引用率看，皮尔斯对该领域的冲击力是如此之大，真不明白现代教科书为何如此冷待皮尔斯。原因也许在于，比起明斯基和乔姆斯基，皮尔斯的批评“麻烦”更大。很多学者试图回应他的批评，但几乎没有任何回应能像他原来的批评那样有力和值得一读。

皮尔斯一生硕果累累，他开发了脉冲编码调制(pulse code modulation, PCM)，一种与当今WAVE 文档格式紧密相关的语音编码方法，而WAVE 是一种在个人计算机上储存音频文件的流行格式。此外，皮尔斯在真空管领域的研究亦成就卓著，但他又带领团队发明了晶体管，使真空管很快消亡。皮尔斯的研究工作也涵盖卫星领域，后来他作为贝尔实验室的研究副总裁，在把卫星研究转化成商业应用上发挥了关键作用，成功开发了Telstar 1，这是卫星首次在电信领域的商业应用。

总之，皮尔斯是一位具有非凡成就的顶级技术执行官。与他争辩的另一方根本无法与他相提并论，其中包括一些可能会被拒授终身教职的初级教职人员。这是一场不公平的论战。但即便如此，也没有理由忽视他对领域的贡献，哪怕这些贡献给我们带来诸多“麻烦”。

ALPAC 报告与“语音识别向何处去”都非常值得一读。网上很容易找到 ALPAC 报告的原文，但其篇幅较长。如果读者时间有限，建议先从阅读“语音识别向何处去”开始，因为这篇通讯言简意赅，观点明晰。短短两页的通讯基本上论及两条批评意见：

1. 系统评测：皮尔斯反对用演示来评测系统，也反对现今仍流行的各种系统评测方法。“即使给出了统计数据，语音识别的成功与否还是很难测量。总体而言……当……时对于……系统可以达到 95% 的准确率。在……情况下，性能会急遽下降。很难鉴定这种性能的语音识别系统能否成为实用的、经济合理的应用产品。”
2. 模式匹配：皮尔斯反对现今仍流行的模式匹配技术（如机器学习和语音识别），斥之为巧妙的欺

骗：“与科学相比由于更容易取巧而更快成功”。

模式识别批判

皮尔斯以魏岑鲍姆(Weizenbaum)开发的伊莉莎(ELIZA) 程序作为案例来解释“巧妙的欺骗”。虽然伊莉莎很明显并不“智能”，但它或许可以通过图灵测试。伊莉莎批判从此成为对那些看上去比实际能力要强的程序的标准批判。维基百科对“伊莉莎效应”的定义如下：

> “在计算机科学中，所谓的伊莉莎效应，指的是下意识地假设计算机与人类的行为相似的一种趋势。从特定形式上看，伊莉莎效应只是指‘人们阅读被计算机串起的符号序列（特别是单词），往往读出了这些符号并不具备的意义’。更一般地，伊莉莎效应描述的是这样一种情形，仅仅依据系统输出，用户就把计算机系统理解为具有‘其软件控制（输出）不可能实现的内在素质和能力’，或者，‘假设（输出）反映了比实际更大的因果关系’。无论是在特定还是一般形式上，甚至当系统的用户已经知道系统产生的输出是预定不变的，伊莉莎效应都会显著出现。从心理学观点来看，伊莉莎效应源于一种微妙的认知失调，一方面，用户意识到程序编制的局限性，另一方面，他们对程序的输出结果依然盲信。伊莉莎效应的发现是人工智能的一个重要进展，说明利用社交工程原理，而不是显式编程，也可以通过图灵测试。”

魏岑鲍姆在意识到他的伊莉莎程序让公众如此信服后，他自己反而成为人工智能的强烈反对者。以下是从他的著作《难以理解的程序》(Incomprehensible Programs) 中的一个章节摘录的：

> “这两个程序(MACSYMA和DENDRAL) 与其他大多数人工智能程序完全不同，它们牢牢建立在深厚的理论之上……计算机当然还有其他许多重要的、成功的应用。例如，计算机可以操控整个石油精炼厂的流程，可以导航飞船以及监测并在很大程度上操控飞船内的环境，以便宇航员执行任务。这些程序依赖于数学控制理论和牢固确立的物理理论。这种以理论为基础的程序具有极其重要的优势，一旦程序走偏，监测人员就能发现它们的性能不符合理论的要求，从而可以用理论帮助诊断失败的原因。
>
> 但是， 大多数现有的程序……不是以理论为基础的……它们多是探索式的……采用的是在多数预见情况下显得‘可行’的策略……我自己的程序伊莉莎正是这种类型。伍诺格拉德(Winograd)的语言理解系统也是……纽厄尔(Newell) 和西蒙的GPS20 也是如此。”

魏岑鲍姆继续争辩道，程序理应易于理解，并建立在坚实的理论基础之上，这种观点皮尔斯想必也会同意。

皮尔斯关于“巧妙的欺骗”的提法批评了包括人工智能、语音识别以及模式识别（也包括大部分现代机器学习）的很多领域用演示来验证系统的做法。

> “前述讨论适用于模式识别的各个领域，其应用作为练习留给读者”。

模式识别有其优缺点。优点是，模式识别可以巧妙应对许多科学难题，在实际应用中取得进步。但是这一优势同时也是其缺点。短期的取巧分散了领域的精力，无法顾及真正有意义的长远目标。

很多工程任务与语音合成一样有两类研究：一类是实用的工程方法（例如衔接合成和磁带拼接），另一类是雄心勃勃的科学计划（如模拟人类发音的合成）。一般而言，实用的方法更有可能在短期内产生较好的结果，但学术界也激励更有前途的科学路线。对于尚未解决的重大科学问题，如果我们直接研究它们，而不是投机取巧，我们会有更好的机会取得进展。话虽这么说，如果你在工业界领导一个语音合成

产品，为了在预算内按时按质交付产品，采用任何工程手段和技巧都是题中应有之义。

回应

针对“语音识别向何处去”曾有很多回应，但是多数回应都没能有效应对上面提到的两条主要批评意见：

1. 目前在论文发表时所要求的系统评测方法究竟有何意义？
2. 与科学相比，模式匹配的意义何在？

罗(Roe)和威尔彭(Wilpon)争辩说，在“语音识别向何处去”提出后的25 年中，领域的发展已经把所谓“无用”的努力演变为商用现实。他们的文章开头介绍了隐式马尔可夫模型等流行方法，这些方法基于皮尔斯所反对的模式匹配技术。接着提到目前常用的评测方法。评测旨在展示模式匹配技术的有效性，然而评测带来的结论正如皮尔斯归纳描述的那样：“难以度量”。

> “在实验室条件下，语言识别器对于声音的模式匹配相当准确。然而，在‘真实世界’的条件下，错误率会高出很多”。

ALPAC报告

相当长的ALPAC 报告提出了很多反对意见，其中许多批评意见令人尴尬，也很难回应。报告的结论部分提到一些好消息：

> “如今仍有理论语言学家对实证研究或计算都不感兴趣，也有应用语言学家对十年来的理论进展无动于衷，对计算机也很木讷。但是，与以往任何时候相比，都有更多的语言学家尝试把微妙的语言理论与更丰富的数据相结合，他们中几乎所有人，无论在哪个国家，都渴望计算机的支持。前一代人需要一辈子做的工作（譬如建立对照语库、词汇表、浅层文法），如今借助计算机几个星期即可完成（下一年大概只需要几天）。在对于作为人类交流工具的自然语言的理解方面，人类迈出了万里长征的第一步。”

但好消息随后紧接着就是不那么好的消息 ：

> “ 但是，我们还没有简单易用并广为人知的计算机处理语言数据的好方法。”

作为回应，斯蒂德曼(Steedman)将我们的研究领域与物理学领域做了对比。他指出物理界并没有被类似于ALPAC 的报告所困扰：“没人去告诉周围的物理学家该做什么。”斯蒂德曼建议，如果我们更自律，并避免在公共场合过度渲染，我们的领域也许会处于更好的状态。

我们其实没必要羡慕物理学领域的状态，以此排斥ALPAC报告。斯蒂德曼的回应不仅没有解决问题，而且事实上，物理学在学术界根本就不处于一个令人羡慕的位置。曾经有一段时间，物理学确实处于相对良好的状态，但那是很久以前的事情了。物理学的冬天已经持续太久，以至于许多人离开了物理学领域。曾经的物理学家们对许多领域做出了贡献，包括我们领域的几个方向，例如机器翻译和机器学习等。至于过度渲染，物理学不比我们少。

甚至连ALPAC 报告也指出，计算语言学比物理学有许多优势：

> “我们看到计算机为语言学家带来了一系列的挑战、视角和机会。我们相信，这些可与粒子物理面临

> 的挑战、问题和视界相当。毫无疑问，语言的重要性不亚于任何其他现象。而且计算语言学所需要的工具成本，比起需要数十亿伏加速器的粒子物理学少得多了。”

哈钦斯(Hutchins)在ALPAC报告30 周年纪念时在《机器翻译国际新闻》(MT News International) 中题目为《ALPAC ：著名（抑或臭名昭著）的报告》的文章中，总结道：

> “ALPAC 对机器翻译持怀疑态度是有一定道理的：当时机器翻译的质量无疑非常糟糕，似乎确实没有正当理由获得那么多的资助。报告中也正确地指出需要研制计算机辅助翻译，并强调计算语言学需要更多的基础研究。然而，需要指责的是……”

哈钦斯继而批评ALPAC 报告的观点太过以美国为中心，机器翻译问题本应在更广阔的全球语境中来考虑。既然基调如此严肃，他对以美国为中心的批评就显得相对单薄。如果从美国角度看机器翻译技术质量不好，费用昂贵，难道换一个角度就会对他国合适？

事实上，ALPAC 报告之所以被认为臭名昭著，是因为它的怀疑论直接导致了机器翻译的资金寒冬，尤其是在美国方面。然而，报告（第34 页）实际上建议在两个不同方向上增加经费开支：

1. 对于语言学和计算语言学的长期的基础学术研究，以及
2. 对于实用的、可以短期奏效的提高翻译质量的工作。

第一类基础研究应该以其科学价值为基础，经过同行评议，而评估第二类应用程序应该着重于实用的指标：速度、成本和质量。

皮尔斯的这两个建议凸显出他的两个不同侧面，正由于这种两面性使得皮尔斯能够同时认同乔姆斯基和香农那样两种不同的立场。一方面，皮尔斯是基础科学的坚定支持者。皮尔斯反对任何将科学扭曲成其他东西（例如应用程序）的企图，以及试图以误导性演示和盲目的指标（如今天所例行的各种评估办法）歪曲科学的发展。另一方面，皮尔斯也有实用的一面，他在语音编码、真空管、晶体管和通信卫星等领域所取得的非凡成就就是证明。他是应用型工作强有力的支持者，但所用的规则与基础研究完全不同，比如强调从商业案例出发。应用型工作要按应用型工作来评估（基于商业标准），而科学必须按科学的标准来评估（基于同行评审）。

如果皮尔斯今天还活着，他会被学术界的现状深深困扰。太多的资金投入到了模式匹配技术和数值评估上，干扰了他认定的作为核心科学问题的学术发展。

从更积极的方面看，皮尔斯的应用一面应该会对谷歌的商业成功留下深刻印象，尤其是在搜索方面。尽管如此，谷歌的边缘业务如语音识别和机器翻译是否可以称作成功，从他的角度应该还有疑问。虽然我们有理由对这些领域抱有希望，像皮尔斯这样的怀疑论者会觉得，比起过去的几十年研发的巨额投资，机器翻译和语音识别的应用成就并不相称。作为一个合理的投资回报，现在的语音识别和机器翻译应该产生一个杀手锏级的应用，使得几乎每个人每天都离不开它，就像当年AT & T 发明的电话，或者像微软Windows 系统或谷歌搜索一样。谷歌在搜索方面的核心业务已经实现了这个理想，也许有一天他们的语音和翻译等边缘业务也能最终达到这一目标。

皮尔斯能给今天的我们提供什么？迄今为止，该领域已经做得很好，采摘了不少低枝果实。在有很多果实容易采摘的好时光里，我们自然应该充分利用这些机会。但是，如果这些机会逐渐枯竭，我们最好还是遵循皮尔斯的教诲，认真面对核心科学的挑战，而不是继续寻找不复存在的容易采摘的果实。

无视历史注定要重蹈覆辙

在大多数情况下，机器学习、信息检索和语音识别方面的实证复兴派简单地无视PCM 的论辩，虽然在神经网络领域，感知机附加的隐藏层可以看作是对明斯基和帕佩特批评的让步。尽管如此，明斯基和帕佩特对他们所著的《感知机》出版20 年以来领域进展之缓慢深表失望。

> “在准备这一版时，我们本来准备‘把这些理论更新’。但是，当我们发现自本书1969 年第一次出版以来，没有什么有意义的进展，我们认为保留原文更有意义……只需加一个后记即可……这个领域进展如此缓慢的原因之一是，不熟悉领域历史的研究人员继续犯别人以前已经犯过的错误。有些读者听说该领域没有什么进步，可能会感到震惊。难道感知机类的神经网络（新名称叫联接主义）没有成为热烈讨论的主题吗？……当然不是，该领域存在很多令人感兴趣的问题和讨论。可能确实也有些现在的发现也会随着时间逐渐显出重要性。但可以肯定的是，领域的基础概念并没有明显改变。今天令人兴奋的问题似乎与前几轮大同小异……我们的立场依然是当年我们写这本书时的立场：我们相信这个领域的工作是极为重要和丰富的，但我们预计其发展需要一定程度的批判性分析，可那些更富浪漫精神的倡导者却一直不愿意去做这种分析，也许因为连通主义的精神似乎变得与严谨分析南辕北辙。
>
> 多层网络并不比感知机更有能力识别连通性。”

计算语言学课程的缺陷

正如上面明斯基和帕佩特指出的，我们不断犯同样错误的部分原因与我们的教学有关。辩论的一方在当代计算语言学教科书中已被遗忘，不再提及，只能靠下一代人重新认识和复原。当代的计算语言学教科书很少介绍PCM 三位前辈。在汝拉夫斯基(Jurafsky) 和马丁(Martin) 编著的教科书以及曼宁(Manning) 等编著的两套教科书中根本没有提及皮尔斯。三本教科书中只有一本简要提起明斯基对感知机的批评。刚刚进入此领域的学生也许意识不到所谓“相关学习算法”包含了很多当今非常流行的方法，如线性回归和logistic回归。

> “一些其他的梯度下降算法(gradient descent algorithms) 有类似的收敛定理，但是在大多数情况下，收敛只能达到局部最优。……感知机收敛能达到全局最优是因为它们从线性分离机这样一类比较简单的模型中选择分类器。很多重要的问题是线性不可分的，其中最著名的是异或问题。……决策树算法可以学习解决这类问题，而感知机则不能。研究人员在对感知机最初的热情[29]消褪以后，开始意识到这些局限性。其结果是，对感知机及相关学习算法的兴趣很快消褪，此后几十年一直一蹶不振。明斯基和帕佩特的论文《感知机》通常被看作是这类学习算法开始消褪的起点。”

曼宁等人的2008 版教科书中有简短的文献指向明斯基和帕佩特1988 年的论文，称其对感知机有不错的描述，但并未提及他们的尖锐批评：

> “对文中提到但本章未进行细述的算法，感兴趣的读者可以参阅以下文献：神经网络方面的毕夏普(Bishop) 、线性和logistic回归方面的黑斯蒂(Hastie) 等人以及感知机算法方面的明斯基和帕佩特等的论文。”

基于这样的描述，学生可能会得出错误印象，以为明斯基和帕佩特是感知机算法（以及当今流行的线性和logistic 回归相关方法）的支持者。

毕夏普明确指出，明斯基和帕佩特绝不是感知机和神经网络的赞许者，而且把它们认作“不正确的构想”予以排斥。毕夏普把神经网络在实际应用中的普及看作是对明斯基和帕佩特上述批评意见的反击证

明，认为并非如他们所说的那样“没有多少改变”、“多层网络并不比感知机更有能力识别连通性”。

当代教科书应该教给学生认识神经网络这类有用的近似方法的优点和缺点。辩论双方都大有可言。排除任何一方的论证都是对我们的下一代不负责任，尤其是当其中一方的批评是如此的尖锐，用到“不正确的构想”和“没有多少改变”这样的说法。

乔姆斯基比皮尔斯和明斯基在当代教科书中被提及得多一些。曼宁和舒兹(Schütze) 的教科书引用乔姆斯基的论文10次，汝拉夫斯基和马丁的教科书的索引中共有27 处引用乔姆斯基的论文。第一本书中较少引用是因为它专注于一个相对狭窄的话题——统计型自然语言处理。而第二本教科书涉及面广泛得多，包括音韵学和语音。因此，第二本书还引用了乔姆斯基在音韵学方面的工作。

两本教科书都提到乔姆斯基对有限状态方法的批评，以及这些批评在当时对经验主义方法论的抨击效果。但是话题迅速转移到描述这些方法的复兴，而对这一复兴的论辩、动因及其对目前实践和未来的影响的讨论则相对较少。

> “由乔姆斯基1956 年的论文开始的一系列极具影响力的论文中，包括乔姆斯基1957 年的论文以及米勒(Miller) 和乔姆斯基1963 年的论文，乔姆斯基认为，‘有限状态的马尔可夫过程’虽然是可能有用的工程探索，却不可能成为人类语法知识的完整认知模型。当时的这些论辩促使许多语言学家和计算语言学家完全脱离了统计模型。
>
> N 元模型的回归开始于耶利内克(Jelinek)、默瑟(Mercer)、巴尔(Bahl) 等人的工作……”

两本教科书对N 元文法的讨论都是从引用其优缺点开始：

> “但是必须认识到，无论怎样解读，‘一个句子的概率’都是一个完全无用的概念……。”
>
> “任何时候，只要一个语言学家离开本研究组，识别率就会上升。”（弗雷德·耶利内克(Fred Jelinek)，当时他在IBM 语音组，1988）

曼宁和舒兹是以这样的引用开始讨论的：

> “统计的考量对于理解语言的操作与发展至关重要。”
>
> “一个人对合法语句的产生和识别能力不是基于统计近似之类的概念。”

这种正反面观点的引用确实向学生介绍了争议的存在，但却不能真正帮助学生领会这些争议意味着什么。我们应提醒学生，乔姆斯基反对的是如今极其流行的一些有限状态方法，包括N 元文法和隐式马尔可夫模型，因为他相信这些方法无法捕捉远距离的依存关系（例如一致关系的限制条件和wh- 位移现象）。

乔姆斯基的立场直到今天仍然是有争议的，本文审阅者之一的反对意见也佐证了这种争议。我不希望站在这场辩论中的某一方。我只是要求应该教给下一代双方的辩论。对于任一方，都不至于由于我们疏于教授而使他们需要重新“发现”。

计算语言学学生应该接受普通语言学和语音学的培训

为了让进入这行的学生对低枝果实采摘完后的情形做好准备，今天的教育最好向广度发展。学生应该全

面学习语言学的主要分支，如句法、词法、音韵学、语音学、历史语言学以及语言共性。我们目前毕业的计算语言学的学生在一个特定的较窄的子领域具有丰富的知识（如机器学习和统计型机器翻译），但可能没听说过格林伯格共性(Greenberg’s universals)、提升(raising)、等同(equi)、 量词辖域(quantifier scope)、 空缺(gapping)、孤岛条件(island constraints) 等语言学现象。我们应该确保从事共指关系(co-reference) 研究的学生都知道成分统制(c-command) 和指称相异(disjoint reference)。当学生在计算语言学会议上宣讲论文的时候，他们应该已经了解形式语言学(formal linguistics) 对此问题的标准处理。

从事语音识别工作的学生需要了解词汇重音（如文献）。音韵学重音对于下游的语音和声学过程具有各种各样的影响。

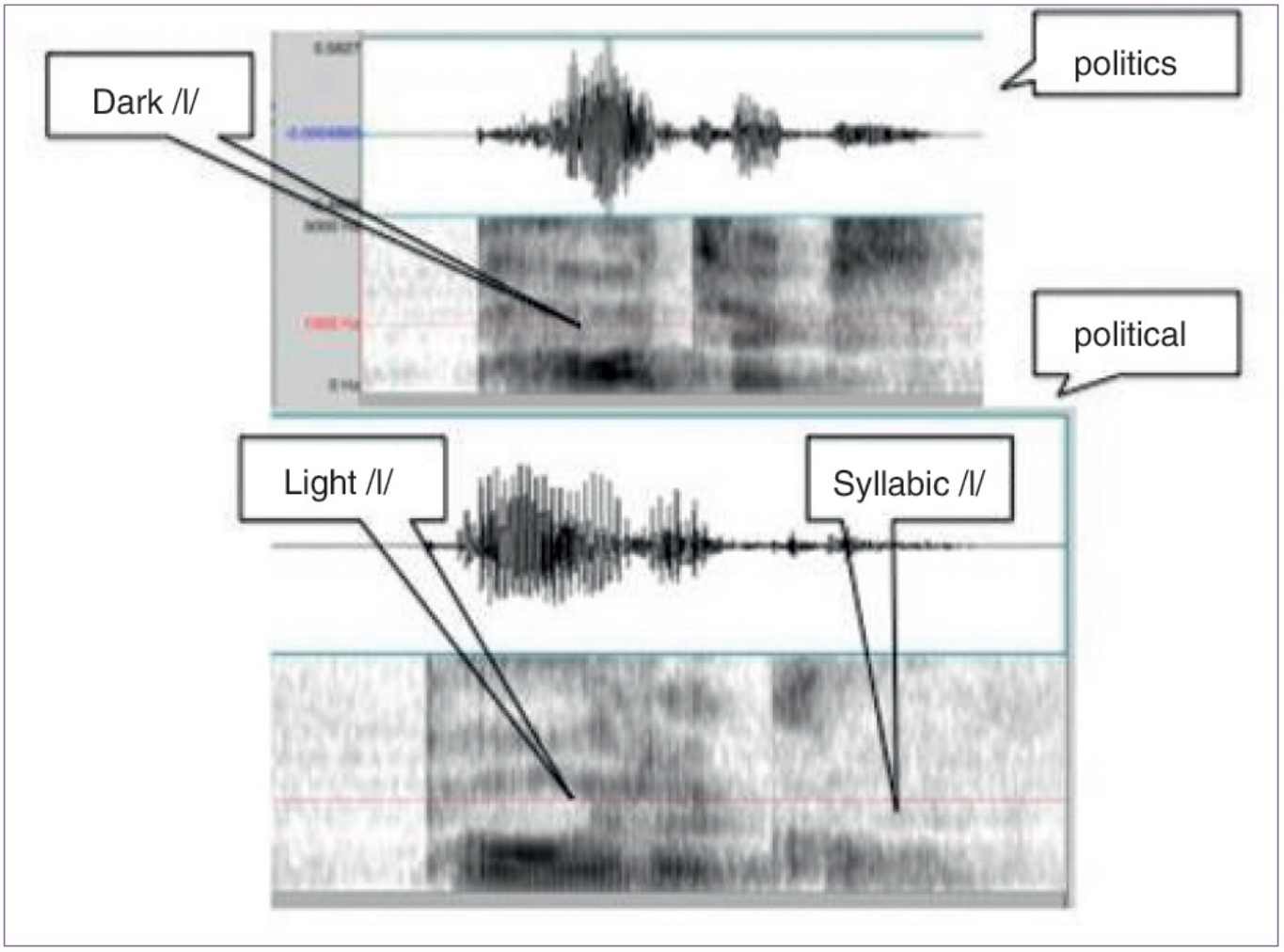


图3 “politics”and“political”的谱图显示有三个/l/同位音。在重音前后出现不同的音位变体。

语音识别目前没有充分利用词汇重音特征是一个不小的遗憾，因为重音是语音信号中较为突出的特性之一。图3 显示了最小对立体 (minimal pair)“ politics”和“political”的波形和谱图。这两个词千差万别，目前的技术着重于语音单位层面的区别：

1. “politics”以 -s 结尾，而“political”以-al 结尾。
2. 与“politics” 不同，“political”的第一个元音是弱化的非重读音节的元音(schwa)。

重音的区别更为突出。在诸多与重音有关的区别中，图3 突出显示了重音前与重音后/l/ 的音位变体之间的区别。另外还有对/t/ 音的影响。“politics”中 /t/是送气音，但在“political”中却是闪音。

目前，在语音单位层面，仍有大量低枝果实可以采摘，但这些工作终有完结之时。我们应该教给语音识别领域的学生有关音韵学和声学语音学的词汇重音知识，以便他们在目前的技术水平超越语音单位层面的瓶颈时依然游刃有余。由于重音存在超过三元音素的远距离依存关系，重音方面的进展需要对目前流行的近似方法的长处与缺陷均有深入的理解。语音识别方面的基础性进展，例如能有效使用重音，很可能要依赖于技术的根本性进步。

结论

学界前辈皮尔斯、乔姆斯基和明斯基曾经严重质疑过当年流行后来复活了的一些经验主义方法。他们的反对意见涉及许多当代流行的方法，包括机器学习（线性分离机）、信息检索（向量空间模型）、语言模型（N 元文法）、语音识别（隐式马尔可夫模型）和条件随机场。

学生们需要学习如何有效地使用流行的近似模型。乔姆斯基指出了N 元文法的缺陷，明斯基分析了线性分离机的局限性。许多局限性很明显（由自身算法设计带来的），但即便如此，对其支持与反对之间的争辩有时仍然非常激烈。有时，其中一方的论点不会被写进教科书，只有等到下一代人去重新发现和复兴这些被遗忘的思想。我们应该鼓励下一代学者充分了解辩论双方的论据，即使他们选择站在一方或另一方。

20 世纪90 年代，当我们复兴经验主义时，我们选择了实用主义的理由来反对我们导师的观点。数据从未如此丰富，我们能拿它做什么呢？我们认为，做简单的事情比什么都不做要好。让我们去采摘一些低枝果实。虽然三元模型不能捕捉到一切语言现象，但它往往比其他方法更有效。捕捉我们可以轻易捕获的一致性事实，要比好高骛远试图捕捉更多语言事实而最终得到更少要好。

这些说辞在20 世纪90 年代有很大的意义，特别是学术界在前一波繁荣期提出了很多不切实际的期望。但是今天的学生在不久的将来可能会面临一系列非常不同的挑战。当大多数低枝果实采摘完毕，他们应该做些什么呢？

具体就机器翻译而言，统计方法的复兴（例如文献）由于实用主义的原因，始于采用有限状态方法。但随着时间的推移，研究人员已经越来越接受使用句法捕捉远距离的依存关系，尤其是当源语与目标语缺乏平行语料库，或者当两种语言具有非常不同的词序的时候（例如，从主谓宾词序的语言（如英语）翻译到以动词收尾的语言（如日语））。展望未来，我们可以预料到机器翻译的研究会越来越多地使用越来越丰富的语言学表达。同样，很快也将有一天，重音将成为语音识别的重要依据。

既然计算语言学教科书不可能涵盖所有这些内容，我们就应该与其他相关科系的同事合作，确保学生能接受到广泛的教育，足以让他们为所有可能的未来做好准备。

选自《中国计算机学会通讯》第9卷第12期。本文译自Linguistics issues in Language Technology, 2011; 6(5) K. Church 的“A Pendulum Swung Too Far”一文。译者：李维（美国网基公司首席科学家。主要研究方向为信息抽取、舆情挖掘等）唐天（美国网基公司首席科学家助理兼助理工程师。主要研究方向为自然语言处理和机器学习）。机器之心授权转载。