# 迎接自然语言处理新时代

李 航
华为诺亚方舟实验室
特邀专栏作家

**关键词：**自然语言理解 自然语言处理

人类的语言具有什么特性？下面是几位最权威学者的看法。

语言是草根现象，它像是维基百科，聚集了数以十万计的人的贡献。当人们要找到更好的表达自己思想方式的时候，就发明了术语、俚语、新说法，其中一部分积累到语言中，这就是我们得到语言的过程。

——史蒂文·平克 (Steven Pinker)

如果语法没有递归结构，那么它将变得不可接受的复杂。因为它有了递归的工具，所以它能够产生无穷多的句子。

——诺姆·乔姆斯基 (Noam Chomsky)

我们通常的概念系统的大部分都具有比喻性。我们的思考方式，我们所经历的，我们每天做的，都与比喻有关。

——乔治·雷可夫 (George Lakoff)

当一个人听到或看到一句话的时候，他使用自己所有的知识和智能去理解。这不仅包括语法，也包括他的词汇知识、上下文知识，更重要的，是对相关事物的理解。

——特里·威诺格拉德 (Terry Winograd)

语言看来是人的认知向外界环境扩展的核心手段。语言的进化也许就是为了扩展我们的认知与外界环境的积极交互。

——安迪·克拉克 (Andy Clark)

总结起来，不完全规则性、递归性、比喻性、知识关联性、交互性是人类语言的主要特点。这些特性密切关联，体现了语言的本质。上述学者对这些语言特性的研究作出了卓越贡献，他们的论述是对这些特性的最佳诠释。

本文从语言的特性出发，讨论为什么让计算机理解人类语言（自然语言）是极其困难的，提出**自然语言处理研究**应该采取的策略。

## 为什么自然语言理解很难?

### 自然语言理解

你说一句话，如何判断别人（或者计算机）是否真正理解了你的意思？这是一个难解的问题。到目前为止，自然语言理解主要有两个定义，一个是基于表示的，一个是基于行为的。对于前者，如果你说“哈利·波特”，别人把它联系到了大脑中的

图1 人通过语言给出命令，机器人若能正确执行，就认为它可以“理解”语言

哈利·波特的概念（表示），那么就认为他理解了你的意思。而对于后者，如果你说“给我拿一杯茶来”，别人真的按你说的做了（行为），就认为他理解了你的意思（图1）。

现在的人工智能研究中，人们开始倾向于采用后者的定义，因为这样更容易评价任务驱动、端到端的语言理解系统的能力。

## 语言的特性

下面结合语言学、认知科学、脑科学的最新研究成果，对语言的主要特性进行介绍。

### 不完全规则性

语言具有一定规范，语言的规范可以用语法来描述，但是，几乎所有的语法规则都存在例外。语法规则中一定有逻辑不一致、功能冗余的现象。正如语言学家爱德华·萨丕尔 (Edward Sapir) 所说，“所有语法都有漏洞 (all grammars leak)”。这是为什么？

其中一个重要原因是，语言不是一个人发明的，甚至不是一组人发明的，而是成千上万人经过成千上万年的时间不断建立起来的，而且在不断演化，这个过程跟人们构建维基百科的过程非常相似。这是认知学家平克等人的观点[1, 2]，也被越来越多的人接受。

语言的基本单元是词汇和语法规则。为了顺畅地交流，需要人们对词汇和语法有基本的共识及准确的使用。另一方面，词汇和语法又不是一成不变的。为了更好地表达自己的思想，人们会不断地去扩展已有词汇和语法规则的使用范围，或者增加新的词汇和语法规则。

语言中不断有大量的新词汇涌现，但其中大部分会逐渐消失，只有真正有生命力的表达才能留存下来。每一个语言的词汇都在不断增加，随着文明的进步，这个趋势会越来越明显。

语法是相对稳定的。在远古时代，语言曾经历过“语法大发明”的时期，后来逐渐趋于成熟。但是即使在现代，语法也不是一成不变的。首先，有一个趋势是语法变得越来越简单。比如，英语中以前说“We shall”、“I shall”，现在逐渐变成“We will”、“I will”。另外，受其他语言影响，语法也会发生变异。比如，非洲美国裔英语（也被称为黑人英语）是受非洲语言影响而形成的一种英语变种，在这个语言中，“I working”、“you working”是正确的说法，笔者猜测可能是受其他语言的影响。

不完全规则性是语言作为人类交流手段而动态发展的必然结果。

### 递归性

现在普遍认为，词汇应该有100万年以上的历史，而语法大概只有7万年左右的历史。而正是在7万年前，智人 (Homo Sapiens)，也就是现在人类的祖先，开始从非洲大陆迁移至欧亚大陆，与此同时开始发明各种语言[1]。

黑猩猩也能使用一些简单的词汇，但我们不认为黑猩猩拥有语言。因为它们不能把词汇组合起来构成句子。组合性、递归性是语言的重要特点。递归的例子如下：“她觉得很好”，“他认为她觉得很好”，“我想他认为她觉得很好”……理论上可以无限扩展。

1956年，语言学家乔姆斯基提出了文法体系，在人类历史上首次用数学模型对语法现象做出严谨的刻画。乔姆斯基特别指出，递归性属于语法的重要特性。只有有了递归这个工具，我们才能够生成无穷多的表达，语言才拥有丰富的表达能力[3]。

### 比喻性

比喻的本质是把表面不相关联的概念，通过它们背后的相似性联系起来。比如微信里的“潜水”。把“潜水”和在微信里“沉默不语”这两个概念联系起来，就是一个比喻。认知科学家雷可夫等认为比喻是语言的重要特性，语言中的发明基本都是基于比喻的[4~6]。

比喻的使用是人类认知能力、语言能力的体现。中文说“开灯”，英语说“turn on the light”，应该始于比喻，开始有一个人或几个人同时发明了这些比喻，后来变成了固定说法，被广泛使用。据观察，

[1] 语言学中，只要有口头语就被认为是“语言”，而不需要有书面语。

一个英语母语的四岁男孩儿，有创意地说出“open the light”（直译就是开灯）。这个例子说明，人天生就有比喻、创造的能力。

比喻是否能被接受并在语言中使用，具有一定的偶然性。一旦比喻变成固定用法，人们就开始习惯性地使用，而不考虑其缘由。比如，中文中所说的“上厕所”、“下厨房”。这些习惯用法都是比喻性的，但是随着时间的推移，已经很难考证当初为什么做出这样的比喻。[2]

比喻也依赖于语言使用的环境与文化。据说，在大多数语言里都有“温暖的爱”这个比喻，如英语中说“warm affection”，在日语中说“暖かい愛”。这些语言都是温带和寒带的语言，热带的语言里就看不到这样的比喻。

## 知识关联性

十几年前，脑科学研究中有一个有趣的发现。当把电极插到猴子的大脑前运动皮质 (pre-motor cortex) 时，有一个脑细胞会在猴子自己吃香蕉和看别人吃香蕉时，同样处于兴奋状态，也就是说对猴子来说这个脑细胞对应着“吃香蕉”的概念。[3]

后来对人脑做类似的实验，但使用功能磁共振。让人实际做和想象做各种动作，比如张嘴和想象张嘴，接球和想象接球。结果发现，对同一动作，实际做和想象做大脑的前运动皮质中发生反应的部位完全一致。

现在一个得到广泛支持的理论认为，对于同一个概念，大脑用固定的脑细胞去记忆，人理解语言的过程，就是激活相关概念的脑细胞，并关联这些概念的过程 [6]。

表示同一个概念的脑细胞，可以通过不同的方式被激活。例如，有一个细胞表示人在喝水，当你看到人在喝水的时候，或者当你从书中读到人在喝水的时候，这个脑细胞同样会被激活。这也能解释为什么我们在读小说的时候常常有身临其境的感觉。

每个人把自己经历的事件进行编码，存储记忆在脑细胞中，在与外界的交互中这些脑细胞被激活，相关的记忆被唤醒。所以，不同人对同样的语言会有不同的理解，因为他们的经历不同。但也有许多共性，因为大家在交流过程中，相互激活对方脑中的表示相同内容的细胞。

发明比喻的时候，大脑中表示两个不同概念的部位都开始兴奋，相关的脑细胞之间产生新的连接，概念之间产生关联，这个过程被称为神经结合 (neural binding)，是现在脑科学研究的重要课题 [6]。

语言的理解实际上动用了大脑中所有的相关知识，是一个非常复杂的过程。这一点在计算机学家威诺格拉德开发的著名的对话系统 SHRDLU 中也有充分体现 [7]。

## 交互性

语言作为人类交流的工具，其重要特点就是交互。哲学家克拉克等人认为，与环境的交互是人或者动物作为智能体存在的必要条件，或者说，离开了与环境的交互，智能就无从谈起 [8]。

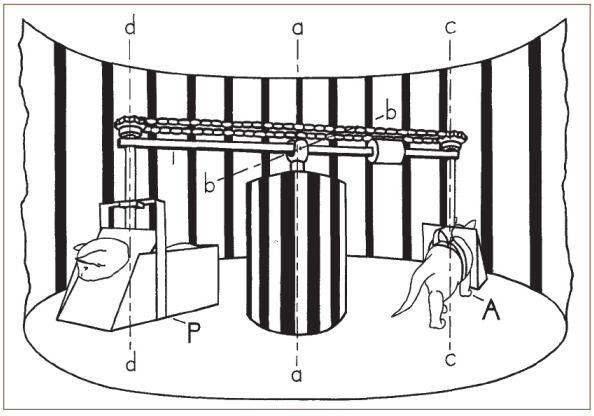


图2　主动猫与被动猫的实验

脑科学家赫尔德 (Richard Held) 和海恩 (Alan Hein) 的实验能够很好地说明与环境的交互对智能体的重要性 [9]。实验对象是一对刚出生的孪生猫，把其中一只当作“主动猫”，另一只当作“被动猫”。白天把它们放到转马上，主动猫脚能着地，可以行走；被

[2] 互联网上有许多关于“上厕所”、“下厨房”语源的讨论。

[3] 猴子和人的运动都是由小脑控制，但大脑的前运动皮质也与运动有关。

动猫被放在篮子里，不能行走。主动猫走动时，转马被带动旋转，这时被动猫也跟着旋转（图2）。晚上把它们放到黑暗处，让它们吃睡。两个月以后，将它们放出去。主动猫和一般的猫没有什么不同，可以正常行走，但被动猫已经失去了行走的能力，走路时要么撞墙，要么跌倒。赫尔德和海恩对10对孪生猫做同样的实验，得到同样的结论。

以上实验说明，对人或者动物来说，虽然拥有先天能力，但在成长的过程中如果不能在与环境的交互中使用，该能力也会丧失。这一点，语言能力也一样。当狼孩被发现时，他已不会说话，因为在他的成长阶段没有与人进行语言交互，没有学习语言。

语言的理解需要在与环境（包括社会、文化）的交互中进行，这点可以在外语学习的过程中体会到。在外语使用环境中学习外语，最容易理解，提高也最快。严格地说，语言是不能翻译的，只能解释。语言必须在其环境中学习与使用。

## 自然语言理解的困难

人的语言理解是一个非常复杂的过程，现在科学对其有了非常粗浅的了解，离理解明了所有细节的程度还相差甚远。

同时，让计算机“理解”人类的语言是极其困难的，因为当代计算机和人脑拥有完全不同的架构。在当代计算机上实现不完全规则性和递归性，意味着进行复杂的组合计算；实现比喻性、知识关联性、交互性，意味着进行全局的穷举计算。是否可行，仍存在巨大疑问。实现能像人一样理解语言的计算机，需要有全新的体系架构，意味着计算机科学发生革命性的进步。

让计算机处理**有限的语言表达**，让它看似很智能，其实不难，只要写出有限的规则就有可能做到。这样的系统做出的演示往往具有一定的欺骗性，让人误以为实现了语言理解。其实一个系统能够理解语言意味着**理论上能够理解无穷多的语言表达**。例如，表1给出了“给我拿一杯茶来”的部分同义说法，理论上类似的表达是无穷多的，一个能理解语言的计算机应该能够判断这样的表达都是同一个意思。而这不是一件容易的事情。关键是要让计算机拥有强健的、通用的语言处理能力。

表1 “给我拿一杯茶来”的同义说法

| |
|---|
| 给我拿杯茶来吧。 |
| 你好，能给我一杯茶吗? |
| 我要一杯茶。 |
| 我渴了，那边好像有茶。 |
| 一到这个时间，就想喝茶。 |
| …… |

## 人们的错觉

人们通常认识不到计算机的自然语言理解极具困难这一事实，可能有以下几个原因。

自然语言具有一定的规律。很多人以为只要写一些规则就可以实现自然语言理解系统，这只是看到了一些非常表面的现象。

人脑的信息处理大部分都是在下意识中进行，有人说其比例高达98%。意识进行的是顺序处理，下意识进行的是并行处理。语言处理也一样。也就是说，人脑进行的大量的语言处理，我们自己是感受不到的。认为语言理解比较简单实际上是我们的错觉。正如彩虹、日出、日落，我们所能直观感受到的，只是现实中发生的很小一部分。

绝大部分人可以在12岁之前几乎无障碍地学会自己的母语，在这个过程中，伴随着大脑的发育，可以在很短的时间内掌握大量的词汇和复杂的语法规则。这个现象是一种奇迹，仍然是认知科学研究的重要课题。

# 自然语言处理的策略

## 自然语言处理

自然语言理解是困难的，但是“自然语言处理”却是可行的。现实中可以让计算机完成一些特定的语言处理任务，比如自动问答、机器翻译、多轮对话，为人们提供帮助，使计算机成为人类的智能助手。现在已部分实现，在可预见的未来可以基本实现，这也是现在自然语言处理研究的目标。

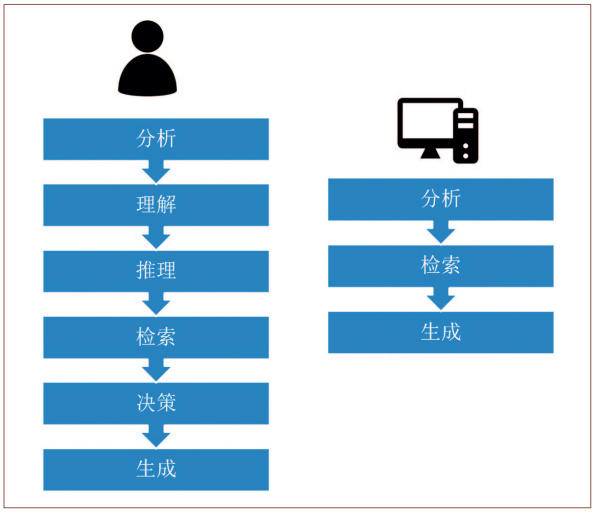


图3 计算机问答处理过程是人的问答处理过程的简化

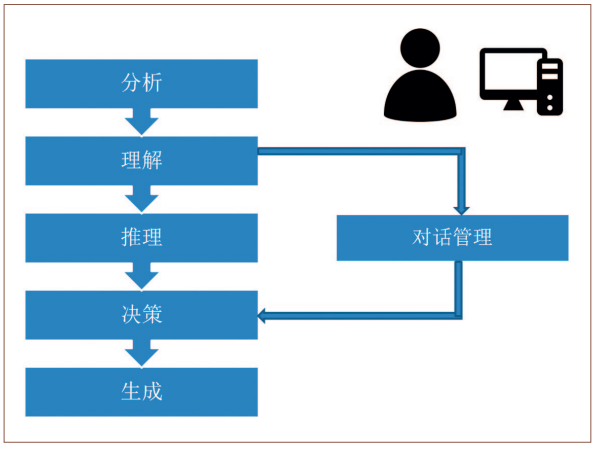


图4 计算机的对话处理过程与人的相似，但适用范围有很大限制

自然语言处理之所以现实可行，主要是因为将人的语言理解过程进行了合理的简化或者限制，而这些简化与限制可以回避自然语言理解中的难题，让计算机表面上像人一样完成语言处理任务。下面以知识问答和多轮对话为例来说明。

人的知识问答可能有这样的处理：得到问题以后，分析问题的内容，理解问题的意思，进行相关的推理，检索相关的知识，决定回答的内容，最后产生回答。现在计算机做知识问答，没有真正的自然语言理解，通常把其中的困难步骤省略简化。计算机的知识问答一般只有以下步骤：分析问题的内容，检索相关的知识，产生回答（见图3）。

人的对话可能有这样的处理：对方发话以后，分析发话的内容，理解发话的意图，进行相关的推理，决定回话的内容，最后产生回话。如果对话是多轮，还有对话管理机制。现在计算机做多轮对话，没有真正的自然语言理解，通常把对话的领域固定，比如订机票、订酒店，并只能在这个领域内进行（见图4）。

## 两大策略

我们认为，自然语言处理可以采用任务驱动与混合模式两大策略。

任务驱动的自然语言处理就是在具体的应用中构建系统。这是现在自然语言处理通常采用的策略，仍可以加强。任务驱动的好处是，可以帮助解决避开自然语言理解之后仍存在的一些问题，而这些问题在实际应用中也相对容易解决。

可以认为自然语言处理经历了三代技术发展演进，第一代基于规则，第二代基于统计，第三代属于现在，基于深度学习。各自有优势和局限。未来的发展方向应该是将这些不同的技术有效地结合起来，即采用混合模式。

### 任务驱动

人工智能系统都遵循这样的规律，我们称作“**人工智能闭环**”（图5）。先有系统，后有用户，然后产生大量数据，机器学习算法可以基于数据构建模型，提高系统的性能，系统性能提高后又能更好地服务于用户，形成一个闭环。人工智能系统可以在这个闭环中不断改进，提升智能水平。自然语言处理也不例外。当任务确定时，就更容易开展基于人工智能闭环的技术开发。

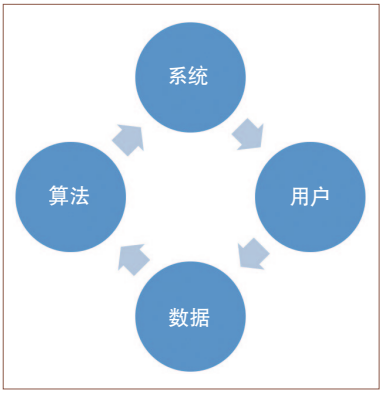


图5 人工智能闭环

### 混合模式

统计方法比起规则方法，能够更好地应对不确

定性。人类的智能，包括语言能力，从数学角度来看，最大的特点就是拥有不确定性。事实证明，统计方法是应对不确定性的最有利工具。

统计方法可以从数据中概括出概率统计规律，构建模型，拥有举一反三的泛化能力。规则方法则不具备这一能力。

深度学习本质也是统计方法，其特点是复杂非线性模型的学习。相比之下，传统的统计方法的模型都是简单的。事实证明，相比传统的统计方法，深度学习有更强的模式学习能力，能够更好地处理复杂的模式识别问题。

规则方法可以有效地利用人给定的知识，而统计方法和深度学习方法，至少是现在，还没有和知识推理有效地结合起来。

统计方法、深度学习方法都依赖于数据。在没有数据或数据稀少的情况下，很难有用武之地。而规则方法，在这种情况下，至少可以派上一定用场。

综上所述，规则、统计（即统计机器学习）、深度学习三种方法都各有优势和局限（见表2）。可以预见，将三者有效地结合，会使人工智能、自然语言处理的水平大幅度提升，这是自然语言处理未来的发展方向。

表2 三种方法的比较

| | 应对不确定性能力 | 泛化能力 | 模式识别能力 | 利用知识程度 | 需要数据程度 |
|---|---|---|---|---|---|
| 规则方法 | 弱 | 弱 | 中 | 大 | 少 |
| （传统）统计方法 | 强 | 强 | 强 | 少 | 大 |
| 深度学习 | 强 | 强 | 极强 | 少 | 极大 |

华为研究团队最近提出了受教式人工智能 (Educated AI, EAI) 的想法，认为这是未来人工智能的范式。其核心思想是，人工智能系统拥有基本的处理以及学习能力，在用户的指导下不断提高智能水平[10]。受教式人工智能采用的就是混合模式，因为人的指导有时是以规则的形式呈现的。

## 自然语言处理新时代

表3总结了现在自然语言处理在各个任务上所能达到的水平，是从不同数据集上得到的实验结果。可以看出，自然语言处理距离人们的期待还有一定的差距，现实中这些任务也只是部分实现了实用化。

表3 现在自然语言处理技术达到的水平

| 任务 | 场景 | 准确率 |
|---|---|---|
| 对话 | 单轮对话 | 80%~90% |
| | 多轮对话 | 60%~70% |
| 自动问答 | 知识问答 | 70%~80% |
| 机器翻译 | 文章翻译 | 70%~80%（由BLEU值推算） |

可以预见，在不远的将来，随着自然语言处理技术的进步，这些性能指标会不断提升。事实上，近年深度学习在自然语言处理的应用，已使机器翻译、单轮对话有了令人惊喜的进步。计算机能够“自如地”进行自然语言处理的时代为期不远。人工智能闭环会推动技术的不断改进，规则、统计、深度学习的结合会产生更强大的技术。现在我们正在进入自然语言处理的一个全新的时代！■

**致谢：**

感谢陈晓博士对本文的评论与建议。

**李 航**

CCF专业会员，*CCCF*特邀专栏作家。华为技术有限公司诺亚方舟实验室主任。主要研究方向为信息检索、自然语言处理、机器学习等。

HangLi.HL@huawei.com

## 参考文献

[1] Pinker S. *The Language Instinct*, 1994.

[2] Pinker S. Linguistics as a Window to Understanding the Brain. *Big Think*, 2013.

[3] Chomsky N. Three models for the description of language [J]. *IRE Transactions on Information Theory*, 1956, 2(3):113-124.

[4] Taylor J. *Linguistic Categorization: Prototypes in Linguistic Theory*, 1996.

[5] Lakoff G, Johnson M. *Metaphors We Live by*, 1980.

[6] Lakoff G. *What Studying the Brain Tells Us About Arts*

*Education*, 2013.

[7] Winograd T. Understanding Natural Language [J]. Cognitive Psychology, 1972, 3(1):1-191.

[8] Clark A. Supersizing the Mind: Embodiment, Action, and Cognitive Extension, 2010.

[9] Held R, Hein A. Movement-Produced Stimulation in Development of Visually Guided Behavior [J]. Journal of Comparative and Physiological Psychology, 1963, 56(5):872-6.

[10]李航, 张宝峰, 霍大伟等. 华为研究的畅想：Educated AI. 中国计算机学会通讯, 2016, 12(1): 62-65.