第十六章 NLP

16.0 NLP 发展史简述

50多年来 NLP 的历史发展可以分为三个浪潮，前两波以理性主义和经验主义的形式出现，为当前的深度学习浪潮铺平了道路。NLP的深层学习革命的主要支柱是：（1）语言嵌入实体的分布式表征，（2）由于嵌入而产生的语义泛化，（3）自然语言的大跨度深序列建模，（4）能够从低到高表示语言层次的分层网络，以及（5）解决许多联合 NLP 问题的端对端深度学习方法。

第一个浪潮：理性主义

在第一个浪潮中，NLP的实验持续了很长一段时间，可以追溯到20世纪50年代。1950年，阿兰·图灵提出了图灵测试，以评估计算机表现出与人类无法区分的智能行为的能力。这项测试是基于人类和计算机之间的自然语言对话，旨在生成类似人类的反应。1954年，George-IBM 实验产出了能够将60多个俄语句子翻译成英语的rrst机器翻译系统。

这些方法是基于这样一种信念，即人类思维中的语言知识是由泛型继承提前进行的，而这种信念，在大约1960年至1980年代后期，占据了NLP的大部分研究中的主导地位。这些方法被称为理性主义方法（Church 2007）。理性主义方法在 NLP 中的主导地位主要是由于诺姆·乔姆斯基（Noam Chomsky）关于先天语言结构的论点被广泛接受以及他对 N-grams 方法的批评（Chomsky 1957）。理性主义者一般假设语言的关键部分在出生时就被硬连接到大脑中，作为人类遗传遗传的一部分，因此他们试图设计手工制作的规则，将知识和推理机制纳入智能 NLP 系统。直到20世纪80年代，最著名的成功的NLP系统，如为模拟 Rogerian psychotherapist 的 ELIZA 系统和为了规则化真实世界信息为规则本体的 MARGIE 系统，都是基于复杂的手写规则。

这一时期恰逢以专家知识工程为特点的早期智能的早期发展，即领域专家根据其所掌握的（非常狭窄的）应用领域的知识设计计算机程序（Nilsson 1982; Winston 1993）。专家们使用符号逻辑规则设计了这些程序，这些规则基于对这些知识的仔细表征和工程。这些以知识为基础的智能系统往往通过检测"Head"或最重要的参数，并就每种特殊情况采取特定的解决办法，而这在解决狭义问题方面往往是有效的。这些“Head”参数由人类专家预先确定，使“tail”参数和案例不受影响。由于缺乏学习能力，他们有必要将解决方案推广到新的情况和领域。这一时期的典型方法是专家系统所提供的证据，这是一个模拟人类专家决策能力的计算机系统。这种系统旨在通过知识推理来解决复杂的问题（Nilsson 1982）。第一个专家系统建立于1970年代，然后在1980年代推广。使用的主要"算法"是以"if-then-else"为形式的推断规则（Jackson 1998）。这些智能系统的主要优点是其在进行逻辑推理方面（有限）能力的透明度和可解释性。像NLP系统，如 ELIZA 和 MARGIE ，一般专家系统在早期使用手工制作的专家知识，这往往是有效的狭隘的问题，虽然推理无法处理不确定性，是普遍存在的实际应用。

同样，语音识别研究和系统设计，这又是另一个长期存在的 NLP 和反智能挑战，在这个理性主义时代，主要基于专家知识工程的范式，如elegantly analyzed in（Church and Mercer 1993）。在1970年代和1980年代初，专家系统的语音识别方法相当流行（Reddy 1976; Zue 1985）。然而，研究人员敏锐地认识到，缺乏从数据中学习和处理推理不确定性的能力，导致了接下来描述的第二波语音识别、NLP和对于文本的人工智能浪潮也走向失败。

第二波浪潮：经验主义

第二波 NLP 浪潮的特点是利用语料库数据以及基于（浅层）机器学习、统计学等来利用这些数据（Manning and Schtze 1999）。由于许多自然语言的结构和理论都被贬低或抛弃，而倾向于数据驱动的方法，这个时代发展的主要方法被称为经验或务实的方法（ChurchandMercer 1993;Church 2014）。NLP 的一个主要会议甚至被命名为“自然语言处理的经验方法（Empirical Methods in Natural Language Processing）（EMNLP）”，最直接地反映了NLP研究人员在那个时代对经验方法的强烈积极情绪。

与理性主义方法相反，经验方法认为人类的思维只是从关联、模式识别和泛化的常规操作开始。丰富的感官输入需要使大脑学习自然语言的详细结构。经验主义盛行于1920年至1960年间，自1990年以来一直在兴起。NLP的早期经验方法主要是开发生成模型，如隐马尔可夫模型 （HMM） （Baum and Petrie 1966）， IBM 翻译模型 （Brown et al. 1993）， 和 head-driven parsing 模型（Collins 1997），以发现大型语料库的规律性。自1990年代后期以来，在各种NLP任务中，歧视性模式已成为事实上的做法。NLP的典型判别模型和方法包括最大熵模型（ratnaparkhi 1997）、支持向量机（Vapnik 1998）、条件随机（Lafferty et al. 2001）、最大相互信息和最小区分器错误（He et al. 2008）还有感知器（Collins 2002）。

在这种经验主义时代中、NLP 与同样的智能方法如语音识别和计算机视觉是平行的。这是在明确的证据表明，学习和感知能力对复杂的智能系统至关重要，但在前一波流行的专家系统中却不存在。例如，当 DARPA 开始对自动驾驶提出重大挑战时，大多数车辆随后依赖于基于知识的智能智能。正如语音识别和NLP 一样，自主驾驶和计算机视觉研究人员意识到基于知识的范式的局限性，因为机器学习需要进行不确定性处理和泛化能力。

在第二波浪潮中，NLP的经验主义和语音识别是基于数据密集型机器学习的，我们现在称之为“shallow”，因为在下一节中描述的第三波浪潮中，数据的多层或“deep”表征通常缺乏抽象结构。在机器学习中，在第一次浪潮中，研究人员不需要考虑构造精确规则，为知识为基础的 NLP 和语音系统。相反，他们把重点放在统计模型（Bishop 2006; Murphy 2012）或作为一个基本引擎的简单的神经网

络（Bishop 1995）。然后，他们使用足够的训练数据进行自动学习或“tune（调整）”系统的参数，使它们能够处理不确定性，并尝试从一个条件泛化到另一个条件，从一个领域泛化到另一个领域。机器学习的关键算法和方法包括EM （期望最大化）、贝叶斯网络、支持向量机、决策树以及神经网络的反向传播算法。

一般来说，基于机器学习的NLP、语音和其他智能系统的性能比早期的基于知识的智能系统要好得多。成功的例子包括语音识别 （Jelinek 1998）， 脸部识别 （Viola and Jones 2004）， 实体识别 （Fei-Fei and Perona 2005）， 手写字体识别 （Plamondon and Srihari 2000）， 以及机器翻译 （Och 2003）。

在语音识别方面，从20世纪80年代初到2010年前后近30年，利用基于 HMM 与高斯混合模型相结合的统计生成模型，以及其推广的各种版本（Baker et al. 2009a，b; Deng and O’Shaughnessy 2003; Rabiner and Juang 1993）的统计生成模式。泛化 HMM 的许多版本都是基于统计和神经网络的隐动态模型（Deng 1998; Bridle et al. 1998; Deng and Yu 2007）。前者采用 EM 和 switching extended Kalman filter 算法学习模型参数（Ma and Deng 2004; Lee et al. 2004），后者采用反向传播（Picone et al. 1999），两者都广泛地利用多个潜在层表示法进行语音分析的生成过程。将这种“深度”生成过程转化为端到端过程的对应方案，导致了深度学习的工业化成功（Deng et al. 2010， 2013; Hinton et al. 2012），从而形成了第三波浪潮的驱动力。

第三波浪潮：深度学习

在第二波浪潮中开发的 NLP 系统，包括语音识别、语言理解和机器翻译，表现得比在第一波浪潮时更好，鲁棒性更高，但它们远远没有达到人的水平，而这留下了很多需求。除了少数例外，NLP的（浅层）机器学习模型通常没有足够的容量来吸收大量的训练数据。此外，学习算法、方法和基础设施也都不够强大。所有这一切都在几年前发生了变化，而这导致了第三波 NLP 浪潮，这股浪潮是由深层机器学习或深度学习的新范式推动的（Bengio 2009; Deng and Yu 2014; LeCun et al. 2015; Goodfellow et al. 2016）。

深度学习起源于人工神经网络，它可以被看作是受生物神经系统启发的细胞类型的级联模型。随着反向传播算法的出现（Rumelhart et al. 1986），90年代对深度神经网络的训练引起了广泛关注。在没有大量训练数据和没有适当的设计和学习范式的情况下，在神经网络训练过程中，学习信号随着层次数（或更严格的信用分配深度）在层层传播时呈指数形式消失，使得调整深层神经网络特别是递归的版本的连接权重变得异常艰难。Hinton 等人（2006）克服了这个问题，使用无人监督的预训练模型来进行学习有用的特征探测器。然后，通过监督学习进一步训练网络，对标记数据进行分类。因此，可以学习使用低维表征的方式来学习高维的表征的分布。这项开创性的工作标志着神经网络的复兴。此后提出和发展了各种网络结构，包括 Deep Belief 网络（Hinton et al.2006）、堆积自编码器（Vincent et al.2010）、深层玻尔兹曼机（Hinton and Salakhutdinov 2012）、深度卷积神经网络（Krizhevsky et al. 2012），深层堆积网络 （Deng et al. 2012），和深层 Q-networks （Mnih et al. 2015）。深度学习自2010年以来已成功地应用于实际智能领域的实际任务，包括语音识别（Yu et al. 2010; Hinton et al. 2012），图像识别（Krizhevsky et al. 2012; He et al. 2016），以及 NLP 绝大多数领域。

其中由于微软公司在工业化上的成功，以及愈来愈高的准确率等迹象，这些2010-2011年语音识别的惊人成功预示着 NLP 的第三波浪潮和人工智能的到来。随着深度学习在语音识别方面取得成功，计算机视觉（Krizhevsky et al. 2012）和机器翻译（Bahdanau et al. 2015）被类似的深度学习范式所取代。特别是，虽然 Bengio 等人在2001的工作，在2011年就开发了强大的神经词嵌入技术（Bengio et al. 2001），但由于大数据的可用性和更快的计算，它直到10多年后才被证明在一个大规模和实际有用的规模上才能够实际有用（Mikolov et al. 2013）。此外，许多其他现实世界的NLP应用，如图像字幕（Karpathy and Fei-Fei 2015; Fang et al. 2015; Gan et al. 2017），视觉问题回答（Fei-Fei and Perona 2016），语音理解系统（Mesnil et al. 2013），网络搜索（Huang et al. 2013b）和推荐系统由于深度学习而取得成功，此外还有许多非NLP任务，包括药物发现和药理学、客户关系管理、推荐系统、手势识别、医学信息、广告投放、医学图像分析、机器人、自动驾驶车辆、纸板和电子游戏（例如 Atari， Go， Poker， and the latest， DOTA2）等。详情请参阅维基上的深度学习领域。

在更多基于文本的应用领域中，机器翻译可能受到深度学习的影响最大。从 NLP 第二波浪潮中发展起来的浅层——统计机器翻译开始看起的话，目前在实际应用中最好的机器翻译系统是基于深神经网络的。例如，谷歌在2016年9月宣布了其转向神经机器翻译的阶段，两个月后微软也发布了类似的声明。Facebook已经进行了大约一年的机器神经网络翻译的转换工作，到2017年8月它已经完全将这个系统部署成功。

在口语理解和对话系统领域，深度学习也正在产生巨大影响。目前流行的技术以多种方式维护和扩展了第二波时代浪潮中发展起来的统计方法。与经验（浅层）机器学习方法一样，深度学习也是基于数据密集型方法，以降低手工制作规则的成本，对噪声环境下的语音识别错误和语言理解错误具有很强的鲁棒性，并利用决策过程和强化学习的力量来设计对话策略，例如（Gasic et al. 2017; Dhingra et al. 2017）。与早期的方法相比，深度神经网络模型和表征方法更强大，它们使端到端学习成为可能。然而，深度学习也没有解决可解释性和领域泛化问题。

将深度学习应用于 NLP 问题方面的最近的两个重要技术突破是序列到序列学习（Sutskevar et al. 2014）和注意力机制建模（Bahdanau et al. 2015），以及最近的 BERT模型（Jacob el al.2018）。序列到序列学习引入了一个强大的学习范式，即使用递归神经网络以端到端的方式进行编码和解码。注意力机制建模最初是为了克服编码一个长序列的难度而开发的，后来的持续发展又扩展了它的能力，提供了两个任意序列的高度可塑对齐能力，而其两个可以同时学习神经网络参数。而 BERT 则是实现了双向建模获取以得到更好的语言表征能力。序列到序列学习和注意力机制的关键概念在基于统计学习和词局部表征的最佳系统上提高了基于分布式单词嵌入的神经机器翻译的性能，而 BERT 更重要的意义是双向获取同一文段的高维意义。在这一成功之后，这些概念也被成功地应用到许多其他与NLP相关的任务中，如图像字幕（Karpathy and Fei-Fei 2015; Devlin et al. 2015）、语音识别（Chorowski et al. 2015）、一次性学习、句法分析、唇读、文本理解、摘要以及问答系统等。撇开他们巨大的经验成功不谈，基于神经网络的深度学习模型往往比早期浪潮中的传统机器学习模型更简单、更容易设计。在许多应用中，在端到端的任务中，模型的所有部分都同时进行深度学习，从特征抽取到预测。导致神经网络模型相对简单的另一个因素是，相同的模型构建成的块（即不同类型的层）通常在许多不同的应用中使用。为多种任务使用相同的构建块，这种方法使得模型更容易迁移到其它任务和数据上。此外，谷歌等公司还开发了软件工具包，以便更快、更有效地实现这些模型。由于以上这些原因，神经网络在数据量大而且基于云的方式上，是更常用的。

尽管深度学习在重塑语音、图像和视频的处理方面被证明是有效的，而且具有它的革命性，但在将深度学习与基于文本的 NLP 相结合方面的有效性并不那么明确，尽管它在一些实用的 NLP 任务中取得了经验上的成功。在语音、图像和视频处理中，深度学习通过直接从原始数据学习规律来解决语义差距问题。然而，在 NLP 中，人们提出了更强的理论和结构化模型，即语音、语法和语义，来提取理解和生成自然语言的基本机制，这些机制与神经网络不那么容易兼容。与语音、图像和视频信号相比，从文本数据中学习的神经表征可以对自然语言提供同样直接的见解，但是这个也不够直接。因此，将神经网络，特别是那些具有复杂层次结构的神经网络应用于 NLP，已成为 NLP 和深度学习社区中最活跃的领域，近年来取得了非常显著的进展（Deng 2016; Manning and Socher 2017;Jacob el al.2018）。

16.1 如何理解序列到序列模型?

16.2 序列到序列模型有什么限制吗?

16.3 如果不采用序列到序列模型，可以考虑用其它模型方法吗?

16.4 如何理解词向量?

16.5 词向量哪家好?

16.6 解释一下注意力机制的原理?

16.7 注意力机制是不是适用于所有场景呢？它的鲁棒性如何？

16.8 怎么将原有的模型加上注意力机制呢？

16.9 通俗地解释一下词法分析是什么？有什么应用场景？

16.10 深度学习中的词法分析有哪些常见模型呢？

16.11 通俗地解释一下知识图谱是什么？有什么应用场景？

16.12 深度学习中的知识图谱有哪些常见模型呢？

16.13 深度学习中的机器翻译有哪些常见模型呢？

16.14 机器翻译的通俗实现以及部署过程是怎样的呢？

16.15 通俗地解释一下文本情感分析是什么？常见的应用场景是？

16.16 最常用的情感分析模型是什么呢？如何快速部署呢？

16.17 通俗地解释一下问答系统？它涵盖哪些领域？常见的应用场景是？

16.18 常见的问答系统模型是什么？如何快速部署呢？

16.19 图像文字生成是什么？它的技术原理是什么？

16.20 常见的图像文字生成模型是什么？

16.21 NLP 的无监督学习发展动态是怎样的？有哪些领域在尝试无监督学习？

16.22 NLP 和强化学习的结合方式是怎样的？有哪些方向在尝试强化学习？

16.23 NLP 和元学习？元学习如何能够和 NLP 结合起来？

16.24 能说一下各自领域最常用且常见的基准模型有哪些吗？