2021-2022学年秋季学期

自然语言处理

Natural Language Processing

授课教师： 胡玥

助　教： 李运鹏

课程信息：

主讲教师

胡玥，中科院信工所，研究员，博导

邮箱：huyue@iie.ac.cn

助教

李运鹏，中科院信工所，2019直博生

邮箱：liyunpeng@iie.ac.cn

课程信息：

课程目标

- 掌握自然语言处理的基本概念、理论、方法
- 掌握正确分析问题、解决问题的思维方式

背景知识

- 概率论、信息论、形式语言与自动机、
- 机器学习（统计、神经网络）
- 基本的语言学知识
- 算法分析基础、编程能力

课程信息：

参考资料：

- 神经网络与深度学习，邱锡鹏
- 近年各顶会相关学术论文
- 基于深度学习的自然语言处理（中译版）车万翔 等 译，机械工业出版社
- 统计自然语言处理（第2版） 宗成庆， 清华大学出版社

课程信息：

课程安排

- 上课时间：周二下午1:30-4:20
- 日期：2021.9.7 – 2022.1.11
- 总学时： 60学时/3学分

考核方式

- 平时作业：60% （实验作业，论文报告）
- 期末考试：40%

自然语言处理
Natural Language Processing

第 1 章　绪 论

授课教师：　胡玥

授课时间：2021.9

课程编码 201M4005H　课程名称 自然语言处理　授课团队名单 胡玥 、李运鹏

概 要

本章主要内容：(一) 自然语言处理的研究背景，发展历史和学派（自然语言处理方法）；(二) 自然语言处理的整体架构，各部分研究内容及发展趋势；(三) 及自然语言处理领域相关的信息：评测，学术会议，国内外研究机构和团队，国际名校的课程，知名的公众号等。

本章教学目的：使学生对自然语言处理领域有整体的认识和概念，包括对各研究内容和发展趋势有所了解；对不同学派处理方法的起源和特点有所了解；和对领域内相关信息有所了解 。

内 容 提 要

1. 自然语言处理与人工智能
2. 自然语言处理发展历史及学派
3. 自然语言处理技术及应用架构
4. 自然语言处理相关信息简介

1. 自然语言处理与人工智能

问题：　什么是当今最热门的技术?

人工智能　（ Artificial Intelligence ）

1. 自然语言处理与人工智能

✧ 人工智能产品

智能客服机器人

阿里小蜜

AI+商业

AI +交通

AI+金融

AI+制造

AI+医疗

AI+教育

AI+农业

1. 自然语言处理与人工智能

人工智能市场未来几年呈现井喷趋势

国际权威调研机构IDC（InternetDataPower）发布最新《中国人工智能软件及应用市场半年度研究报告(2019H2)》报告显示， 2019年中国人工智能整体市场规模将达到60亿美元。到2024年，中国人工智能市场将达到千亿美元规模,人工智能将在未来十年成为科技进步唯一、最大的驱动力

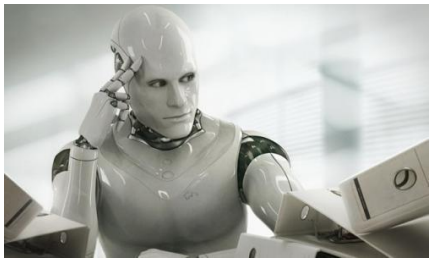
中国人工智能及应用市场研究报告(2019H2)

人工智能将在以下领域产生深远的影响

1. 自然语言处理与人工智能

什么是人工智能?

人工智能 — 建立可智能化处理事物的系统。让机器能够像人类一样完成智能任务（如学习、推理、规划、决策等等）。



1. 自然语言处理与人工智能

人工智能起源

Turing

1950年“计算机科学之父”及“人工智能之父”英国数学家阿兰•图灵在的一篇著名论文《机器会思考吗？》里提出图灵测试：

把一个人和一台计算机分别隔离在两间屋子，然后让屋外的一个提问者对两者进行问答测试。如果提问者无法判断哪边是人，哪边是机器，那就证明计算机已具备人的智能。

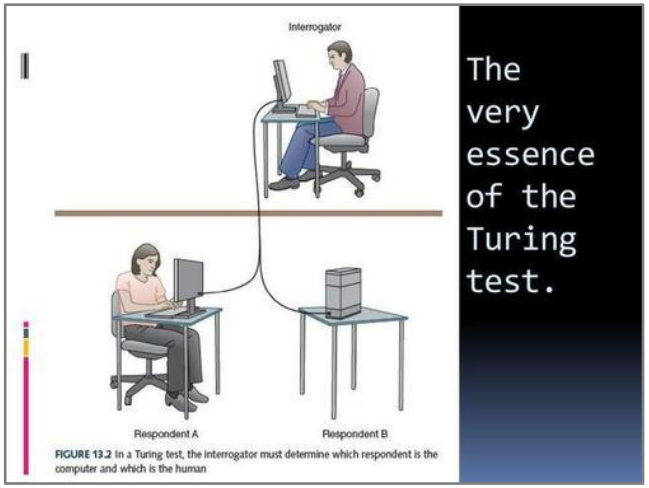


FIGURE 13.2 In a Turing test, the interrogator must determine which respondent is the computer and which is the human

1. 自然语言处理与人工智能

人工智能层次

人工智能主要包括以下三个层次：

第一是运算智能：即记忆、计算的能力；

第二是感知智能：包括听觉、视觉、触觉；

第三认知智能：包括理解、运用**语言的能力**，掌握知识、运用知识的能力，以及在语言和知识基础上的推理能力；

最高层创造智能：人们利用已有的条件，利用一些想象力产生很好的作品或产品。

1. 自然语言处理与人工智能

人工智能基础技术

自然语言处理是人工智能的一个分支，用于分析,理解和生成自然语言，以方便人和计算机设备以及人与人之间的交流比尔.盖茨：“语言理解是人工智能领域皇冠上的明珠”自然语言处理成为人工智能（认知）关键的核心问题

1. 自然语言处理与人工智能

✧ 什么是自然语言

– 自然语言指人类社会发展过程中自然产生是约定俗成的人类语言

– 语言是人类交际的工具，是人类思维的载体

如汉语、英语、日语等，以及人类用与交流的非发声语言，如手语、旗语等。自然语言是相对于人造语言（世界语或计算机的各种程序设计语言）而言的。

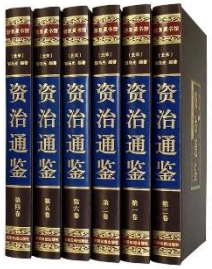
- 形式：口语、书面语、手语
- 语种：汉语、英语、日语、法语...

1. 自然语言处理与人工智能

✧ 什么是自然语言

在整个人类历史上以语言文字形式记载和流传的知识占到知识总量的 80%以上。

据统计计算机应，用于数学计算的仅占 10%，用于过程控制的不到 5%，其余 85%左右都是用于语言文字的信息处理。



1. 自然语言处理与人工智能

全世界正在使用的语言有1900多种不同语言之间结构各有差异

三个不同的语系

- **曲折语: 用词的形态变化表示语法关系，如英语、法语等。**
- **黏着语: 词内有专门表示语法意义的附加成分，词根或词干与附加成分的结合不紧密，如日语。**
- **孤立语(分析语): 形态变化少，语法关系靠词序和虚词表示，如汉语。**

汉语是世界上使用人数最多的语言。大约有13-14亿人

1. 自然语言处理与人工智能

✧ 自然语言处理

自然语言处理（Natural Language Processing，简称NLP）是利用**计算机为工具**，对人类特有的**书面形式**和**口头形式**的自然语言的信息进行各种类型处理和加工的技术。

——冯志伟《自然语言的计机处理》

上海外语教育出版社，1996

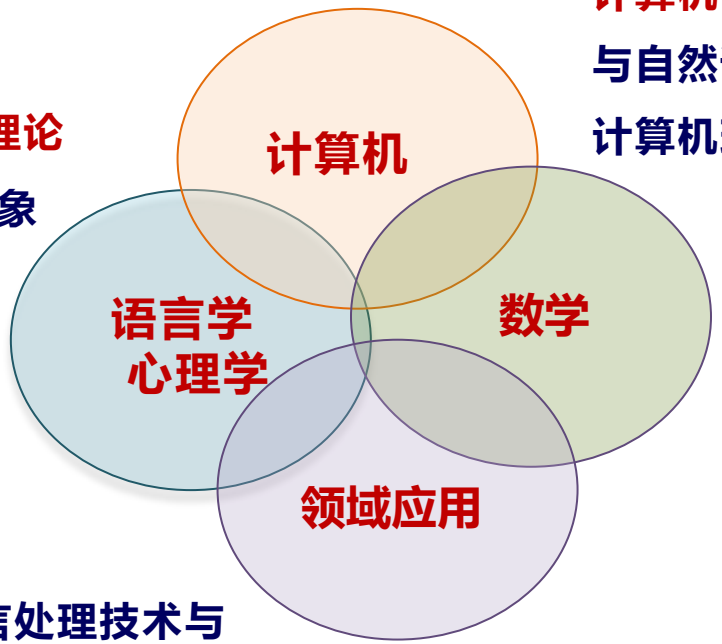
其它名称

- **自然语言理解**(Natural Language Understanding)
- **计算语言学**(CL, Computational Linguistics)
- **人类语言技术**(Human Language Technology)

自然语言处理是人工智能的一个分支，用于分析、理解和生成自然语言，以方便人和计算机设备以及人与人之间的交流

1. 自然语言处理与人工智能

自然语言处理涉及领域



计算机

与自然语言处理相关的
计算机理论与技术

语言学&心理学理论

解释刻画语言现象
的方法及理论

数学

各种用于自然语言处理
的模型及算法

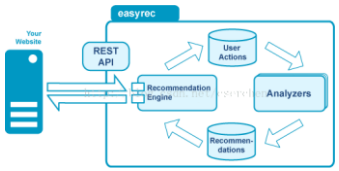
领域应用

利用自然语言处理技术与
领域知识结合的应用系统

认知科学、语言学、逻辑学、应用数学、计算机科学等交叉学科

1. 自然语言处理与人工智能

✧ 自然语言处理应用范围



中国互联网上有87.8%的网页内容是文本表示的

内 容 提 要

1. 自然语言处理与人工智能
2. 自然语言处理发展历史及学派
3. 自然语言处理技术及应用架构
4. 自然语言处理相关信息简介

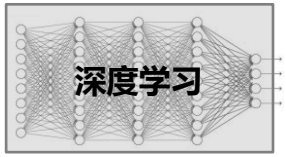
2. 自然语言处理发展历史及学派

理性主义： 1960s – 1980s中期

- 以语言学家 N.Chomsky 为代表。认为人类生成合乎文法的语句的能力是生来具有的,为此他提出一种称为生成句法(Generative Grammar) 的理论
- 理性主义试图去描写人脑中的语言模型（产生模型）：通过一组有限的规则作用于一个有限的词汇集, 生成无限的可接受的、合乎文法的句子
- 技术路线：人工规则方法
- 理论基础：基于乔姆斯基的语言理论

- 以行为心理学家伯尔赫斯·弗雷德里克·斯金纳B. F. Skinner为代表。认为人类语言能力的获得来自于学习,语言是通过不断地实践而“约定俗成”的结果。
- 经验主义试图去刻画真实世界的语言现象
- 技术路线：统计学习方法（数据驱动-从数据中学习的方法）
- 理论基础：基于香农的信息论--将语言事件赋予概率，作为其可信度，由此来判断一个句子是常见的还是罕见的
- 要素：数学基础、统计算法、训练语料

1920s~1950s/1980s中期



2010s~至今

- 仍属于经验主义学派
- 技术路线：神经网络方法（数据驱动-从数据中学习的方法）
- 理论基础：深度学习理论及方法
- 要素：数学基础、深度学习算法、训练语料
- 特点：端到端的解决问题方式

经验主义：

2. 自然语言处理发展历史及学派

理性主义： 1960s – 1980s中期

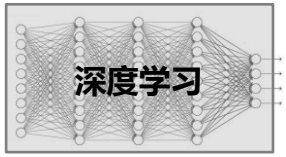
优势： 语言知识的表示直观、灵活；易于表达复杂的语言知识；具有很好的描述能力和生成能力

缺陷： 语言知识的覆盖率低；语言知识的冲突缺乏统一解决机制
劳动强度大，成本昂贵；自然语言是不断发展变化，规则方法应变能力弱

1920s~1950s/1980s中期

优势： 大规模数据提高了语言知识的覆盖率；对自然语言的发展变化应变能力强；统计模型提供了统一的冲突解决机制

缺陷： 不善于表示复杂的、深层次的语言知识；对于数据稀缺的语言（小语种）没有很好的解决办法



2010s~至今

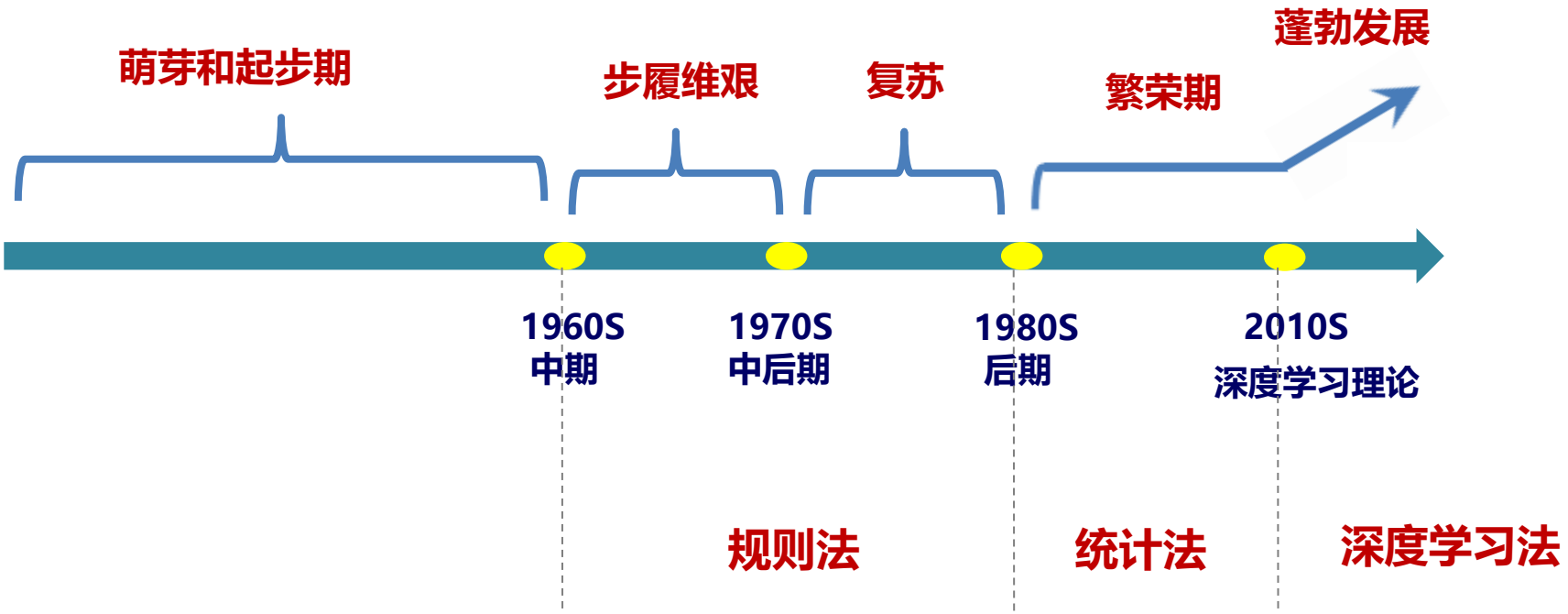
优势： 简化了问题的解决过程，减少错误级联，模型效果显著

缺陷： 解释性差，训练耗时困难，需要大量的领域相关数据

经验主义：

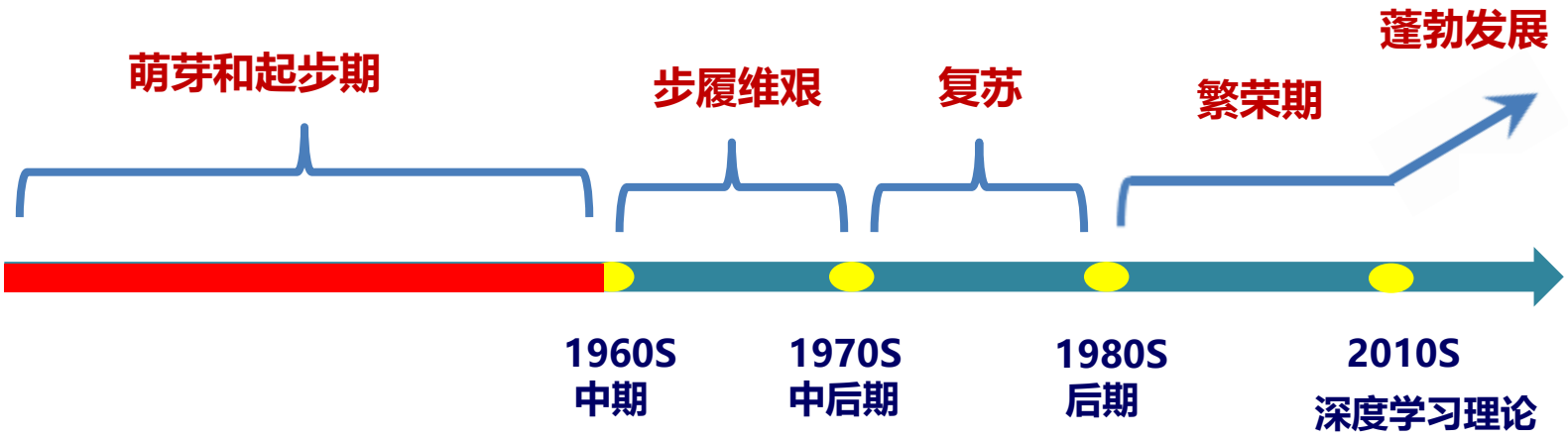
2. 自然语言处理发展历史及学派

作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



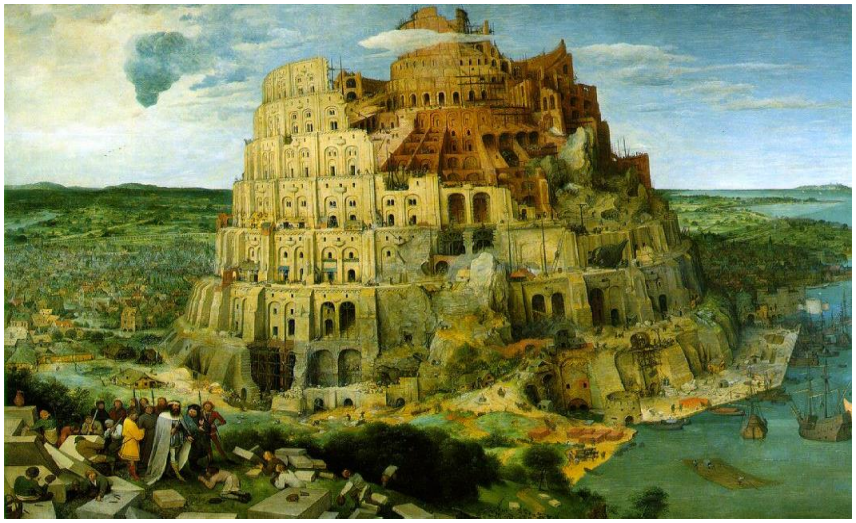
2. 自然语言处理发展历史及学派

作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



1960S 中期之前--萌芽和起步期

巴比塔



人类历史上最早的计算语言学研究就是机器翻译(machine translation)。

圣经《创世纪》中说，古代人类说的原是一种统一的语言，他们曾经想建立一座高达天庭的通天塔，叫做“巴比塔”，此举震惊了上帝，上帝便让不同的人说不同的语言，使人们无法协调工作，结果，巴比塔没有建成，而语言的不同，却成为人们相互交往的极大障碍。

1960S 中期之前--萌芽和起步期

“普遍语言”的运动

在17世纪，一些有识之士提出了采用机器词典来克服语言障碍的想法。

- 笛卡儿（Descartes）和莱布尼兹（Leibniz）都试图在统一的数字代码的基础上来编写词典。
- 在17世纪中叶，贝克（Cave Beck）、基尔施（Athanasius Kircher）和贝希尔（Johann Joachim Becher）等人都出版过这类的词典。由此开展了关于“普遍语言”的运动，一些人试图在逻辑原则和图形符号的基础上，创造出一种无歧义的语言，这样一来，人们就不必再由于误解而产生交际方面的困惑了。

1960S 中期之前--萌芽和起步期

用数学方法研究语言的先驱

- 1847年，俄国数学家B. Buljakovski认为可以用概率论方法来进行语法、词源和语言历史比较的研究。
- 1851年，英国数学家A. De Morgen把词长作为文章风格的一个特征进行统计研究。
- 1894年，瑞士语言学家De Saussure指出，在基本性质方面，语言中的量和量之间的关系，可以用数学公式有规律地表达出来，他在1916年出版的《普通语言学教程》中又指出，语言好比一个几何系统，它可以归结为一些待证的定理。
- 1898年，德国学者F.W. Kaeding统计了德语词汇的在文本中的出现频率，编制了世界上第一部频率词典《德语频率词典》。

1960S 中期之前--萌芽和起步期

1913年，俄罗斯著名数学家A. Markov（马尔可夫）就注意到俄罗斯诗人普希金的叙事长诗《欧根·奥涅金》中语言符号出现概率之间的相互影响，他试图以语言符号的出现概率为实例，来研究随机过程的数学理论，提出了马尔可夫链（Markov Chain）的思想，他的这个开创性的成果后来成为在计算语言学中广为使用的马尔可夫模型（Markov model），是当代计算语言学最重要的理论支柱之一。

Markov

1960S 中期之前--萌芽和起步期

发明家

- 1933年，苏联发明家特洛扬斯基（П．П．ТРОЯНСКИЙ）设计了用机械方法把一种语言翻译为另一种语言的机器，并在同年9月5日登记了他的发明。
- 三十年代之初，亚美尼亚裔的法国工程师阿尔楚尼（G.B. Artsouni）提出了用机器来进行语言翻译的想法，并在1933年7月22日获得了一项“翻译机”的专利，叫做“机械脑”（mechanical brain）。

  阿尔楚尼的原型机于1937年正式展出，引起了法国邮政、电信部门的兴趣。但是，由于不久爆发了第二次世界大战，阿尔楚尼的机械脑无法安装使用。

1960S 中期之前--萌芽和起步期

1936年，Turing 在《论可计算数及其在判定问题中的应用》这篇开创性的论文中，提出著名的“图灵机”(Turing Machine)的数学模型。

1950年 在的一篇著名论文《机器会思考吗？》里提出图灵测试，开启人工智能研究先河。

Turing

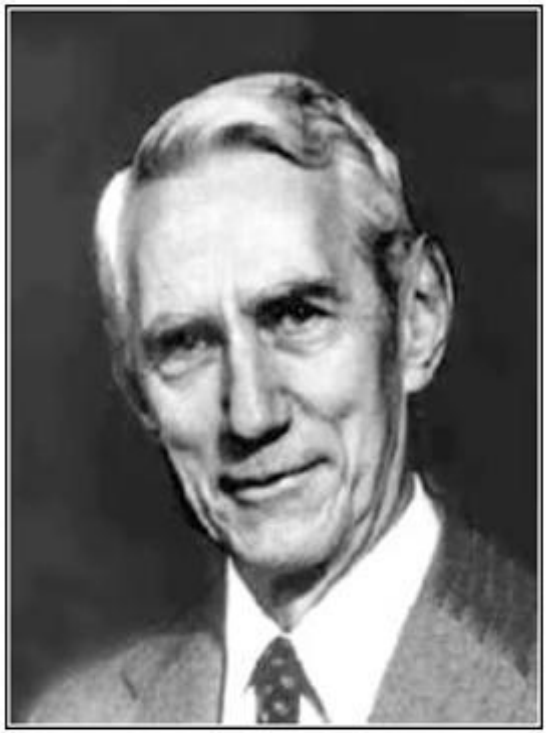
“图灵机”不是一种具体的机器，而是一种抽象的数学模型，可制造一种十分简单但运算能力极强的计算装置，用来计算所有能想象得到的可计算函数。该研究成为现代计算机科学的基础。

1960S 中期之前--萌芽和起步期

1948年，美国学者Shannon（香农）使用离散马尔可夫过程的概率模型来描述语言的自动机。

Shannon的另一个贡献是创立了“信息论”（Information Theory）。他把通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为“噪声信道”（noisy channel）或者“解码”（decoding）。

Shannon还借用热力学的术语“熵”（entropy）来作为测量信道的信息能力或者语言的信息量的一种方法，并且他用概率技术首次测定了英语的熵。



Shannon

1960S 中期之前--萌芽和起步期

机器翻译

1946年 UPenn的J.P.Eckert 和J.W.Mauchly 设计了世界上第一台电子计算机 ENIAC

英国工程师 Andrew Donald Booth 和美国洛克菲勒基金会 (Rockefeller Foundation) 副总裁 W. Weaver提出机器翻译的想法

1949年，韦弗发表了一份以《翻译》为题的备忘录，正式提出了机器翻译问题。他说：“当我阅读一篇用汉语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读时，我是在进行解码。韦弗的卓越思想成为了而后统计机器翻译（Statistic Machine Translation，简称SMT）的理论基础。

韦弗（W.Weaver）

1960S 中期之前--萌芽和起步期

1956年，美国语言学家N. Chomsky（乔姆斯基）从Shannon的工作中吸取了有限状态马尔可夫过程的思想，把有限状态自动机作为一种工具来刻画语言的语法，并把有限状态语言定义为由有限状态语法生成的语言。产生了“形式语言理论”（formal language theory）

采用代数和集合论把形式语言定义为符号的序列.成为计算机科学最重要的理论基石。

Chomsky

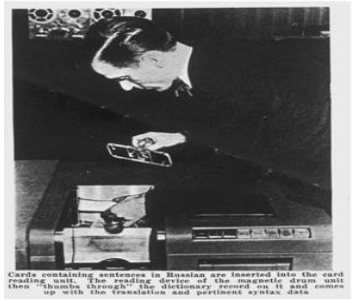
1960S 中期之前--萌芽和起步期

- 美国和英国的学术界对机器翻译产生了浓厚的兴趣，并得到了实业界的支持
- 1954年 Georgetown 大学在 IBM 协助下，用IBM-701计算机实现了世界上第一个 MT 系统，实现俄译英翻译，1954年1月该系统在纽约公开演示
- 在随后10 多年里，MT 研究在国际上出现热潮，一批自然语言人机接口系统和对话系统相继出现

Punched card input

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

Data input

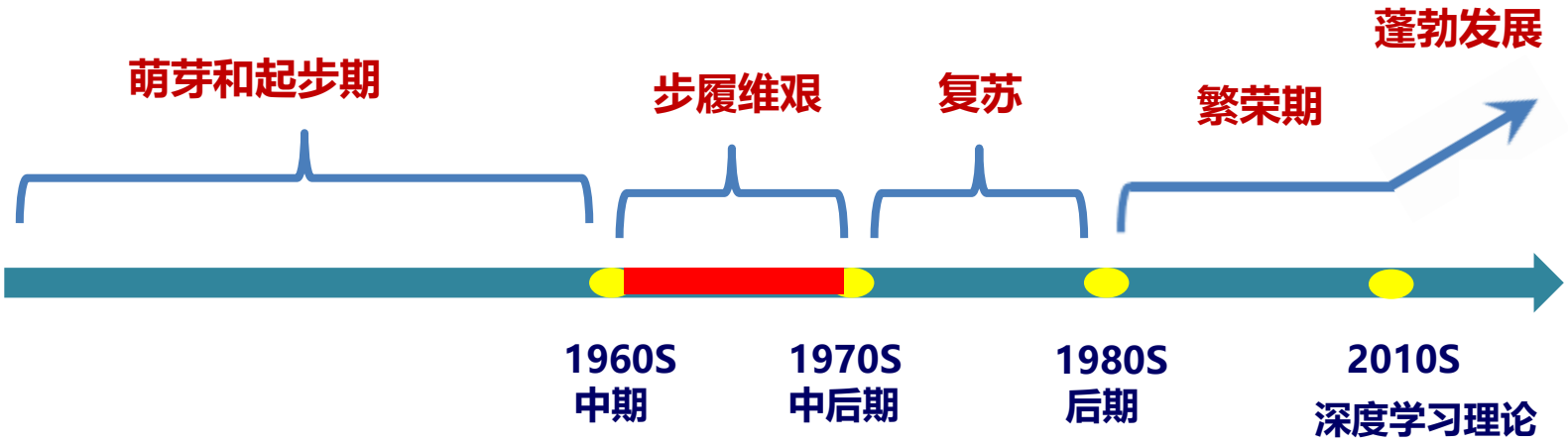


Cards containing sentences in Russian are inserted into the card reader. The machine "looks" at the words, one at a time, and then "thumbs through" the dictionary record on it and comes up with the translation and pertinent syntax data

在此期间随着机器翻译研究的进展，各种自然语言处理技术应运而生，并逐渐发展壮大，形成了这一语言学与计算机技术相结合的新兴学科。

2. 自然语言处理发展历史及学派

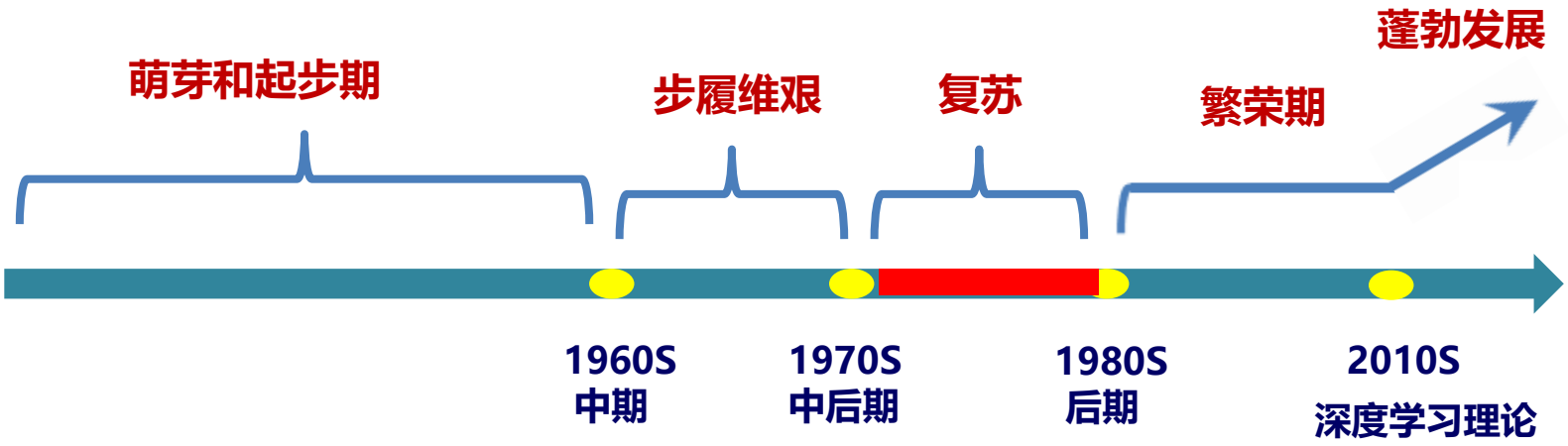
作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



1964年，美国科学院成立了语言自动处理谘询委员会（简称ALPAC委员会），调查机器翻译的研究情况，并于1966年11月公布了一个题为《语言与机器》的报告，简称ALPAC报告。在ALPAC报告的影响下，许多国家的机器翻译研究低潮，，出现了空前萧条的局面。

2. 自然语言处理发展历史及学派

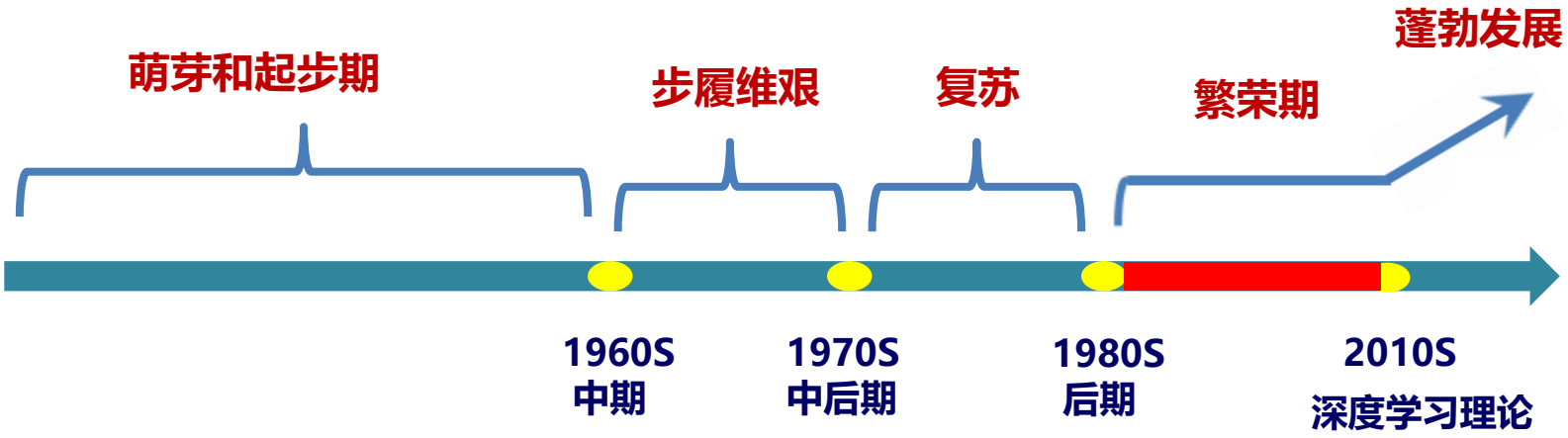
作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



贾里尼克和他领导 的IBM华生研究中心的研究人员以及卡内基梅隆大学的Baker等二支队伍，在统计方法语音识别算法的研制中取得成功：“隐马尔可夫模型”（Hidden Markov Model）和“噪声信道与解码模型”（Noisy Channel Model and Decoding Model）。

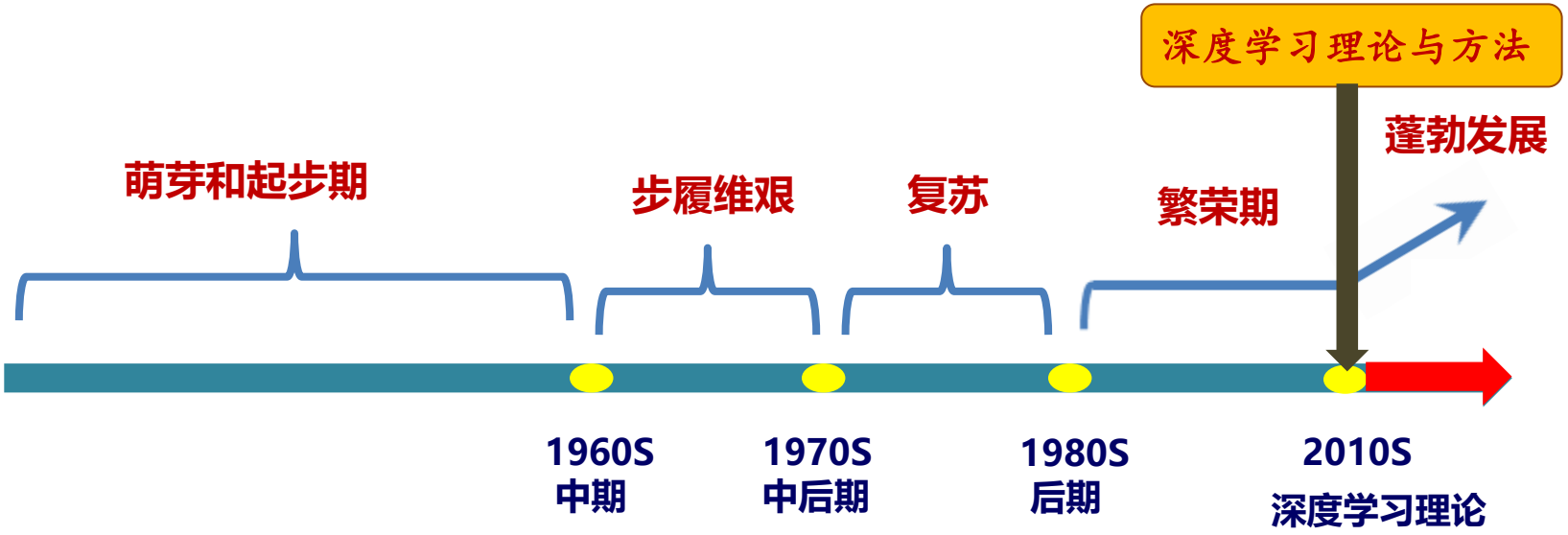
2. 自然语言处理发展历史及学派

作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



1993年7月在日本神户召开的第四届机器翻译高层会议（MT Summit IV）上，英国著名学者J. Hutchins在他的特约报告中指出，机器翻译的发展进入了一个新纪元。随着机器翻译新纪元的开始，计算语言学进入了它的繁荣期。

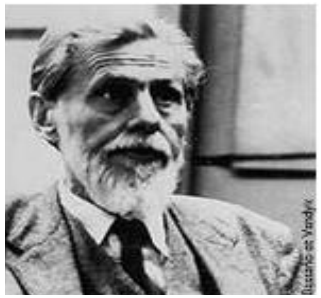
2010 S 进入了蓬勃发展（深度学习）时期



深度学习的发展历史

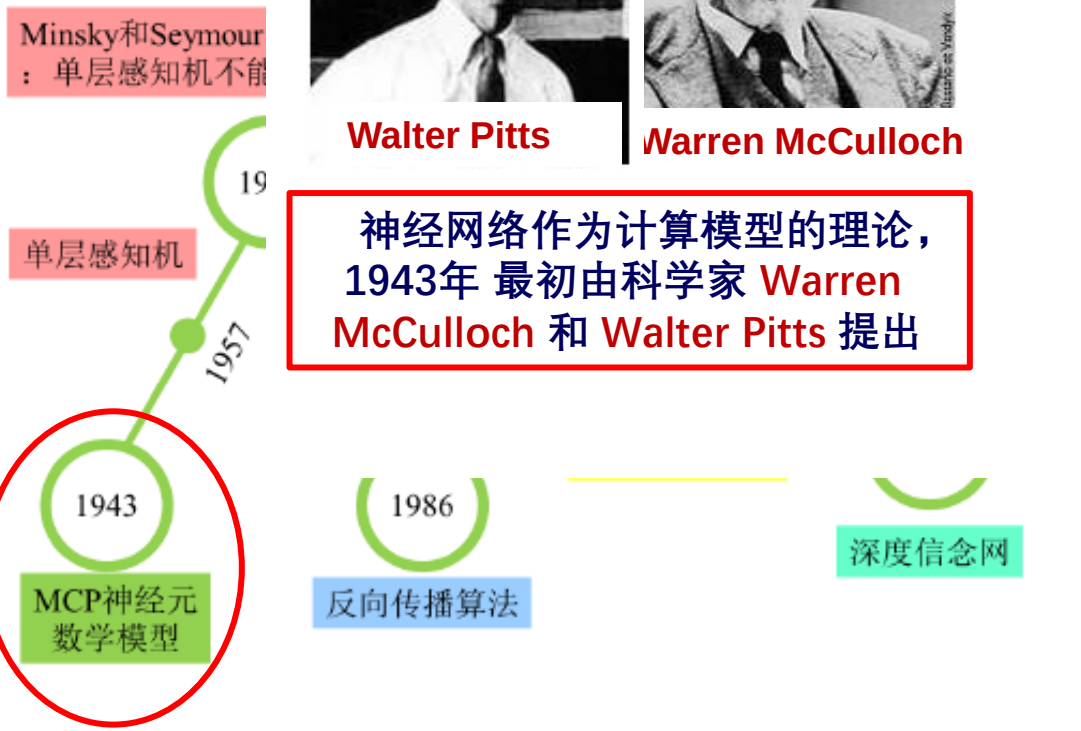
第一代神经网络

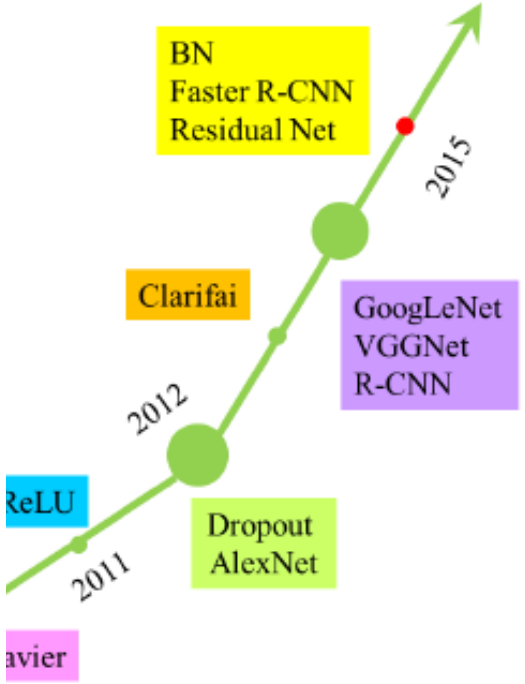
Walter Pitts



Warren McCulloch

神经网络作为计算模型的理论，1943年 最初由科学家 Warren McCulloch 和 Walter Pitts 提出



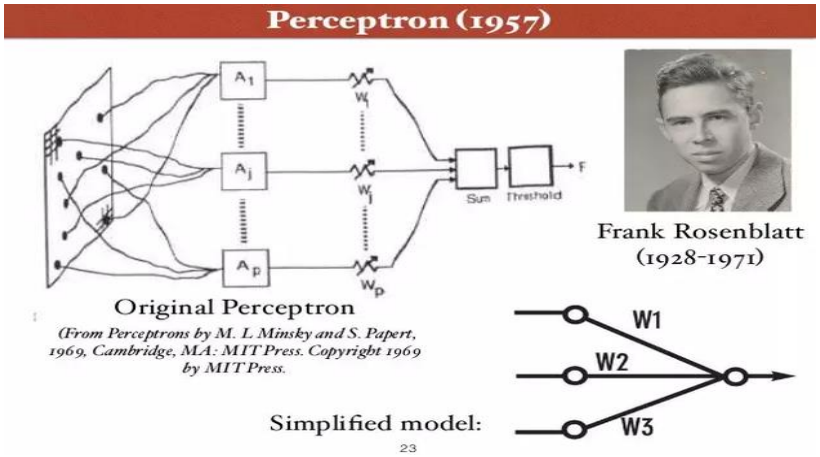


神经网络的开山之作：A logical calculus of the ideas immanent in nervous activity

深度学习的发展历史

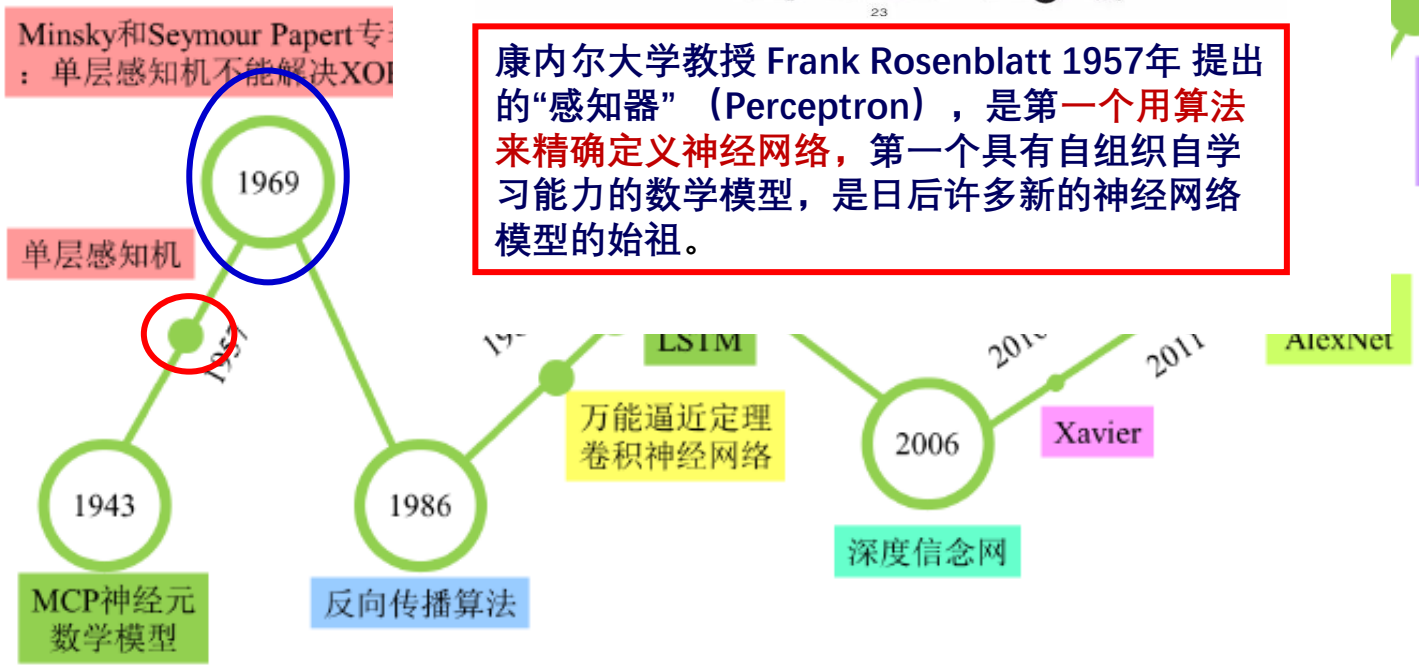
第一代

Marvin Minsky



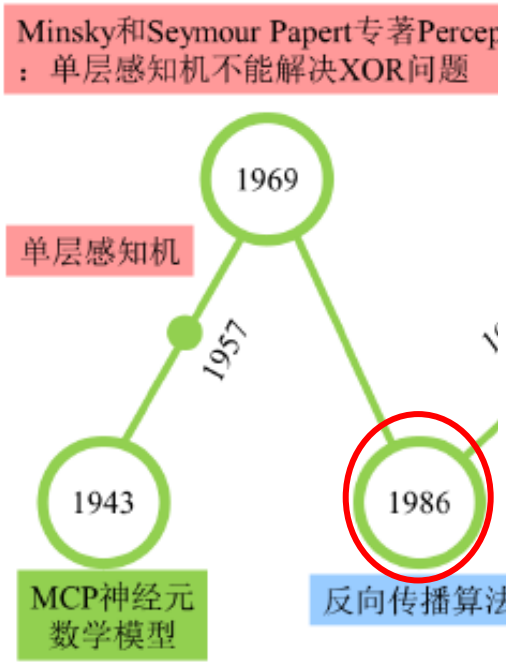
Frank Rosenblatt
(1928-1971)

康内尔大学教授 Frank Rosenblatt 1957年 提出的“感知器”（Perceptron），是第一个用算法来精确定义神经网络，第一个具有自组织自学习能力的数学模型，是日后许多新的神经网络模型的始祖。



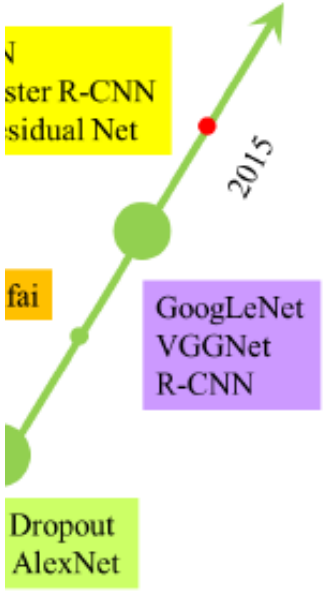
深度学习的发展历史

第二代神经网络



1986年 七月，Hinton 和 David Rumelhart 合作在自然杂志上发表论文，“Learning Representations by Back-propagating errors”，第一次系统简洁地阐述反向传播算法在神经网络模型上的应用。神经网络的研究开始复苏

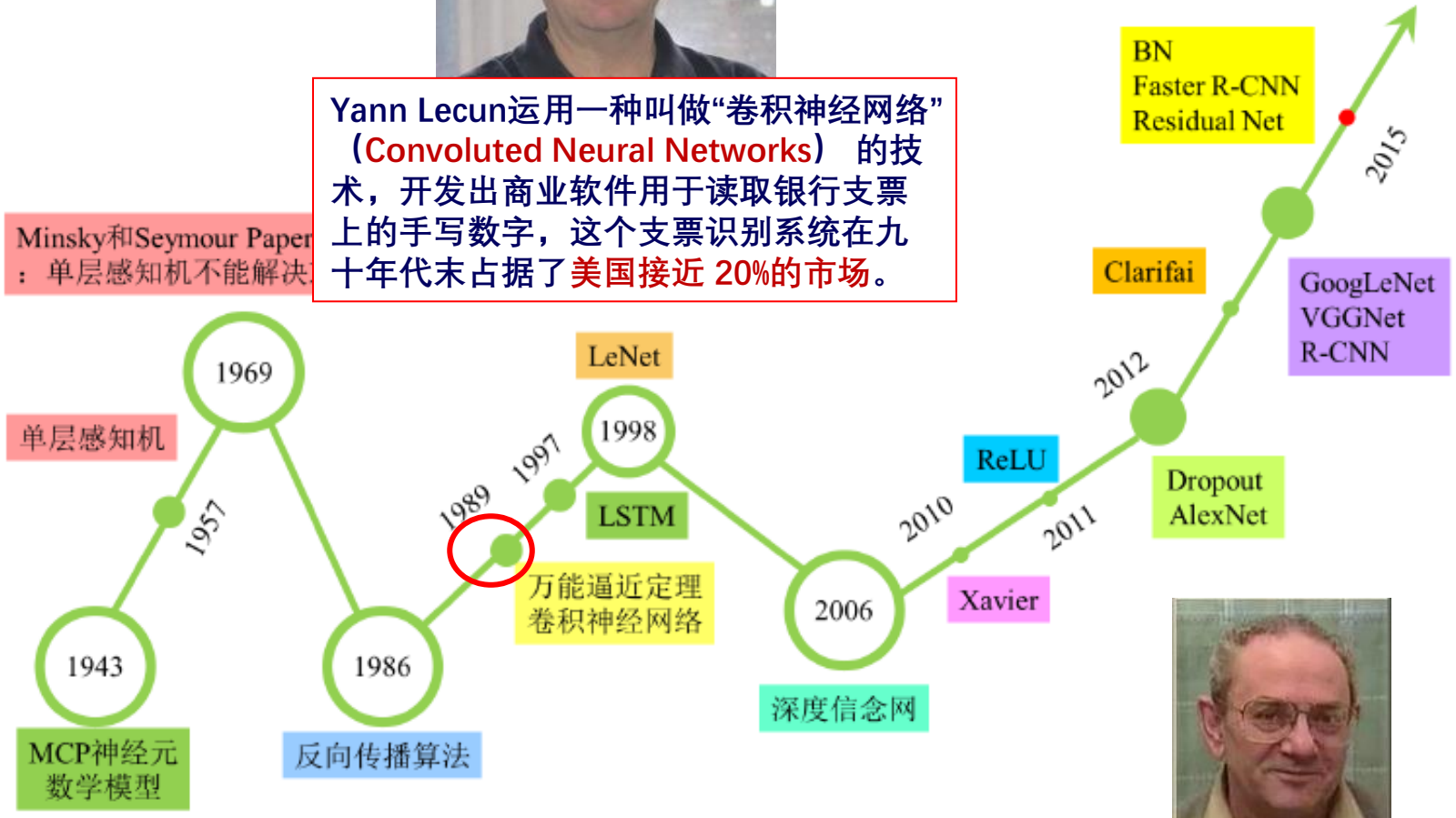
深度信念网



深度学习的发展历史

第二代神经网络

Yann Lecun运用一种叫做“卷积神经网络”（Convoluted Neural Networks） 的技术，开发出商业软件用于读取银行支票上的手写数字，这个支票识别系统在九十年代末占据了美国接近 20%的市场。



Vladmir Vapnik 提出 支持向量机（Support Vector Machine） 的算法。

深度学习的发展历史

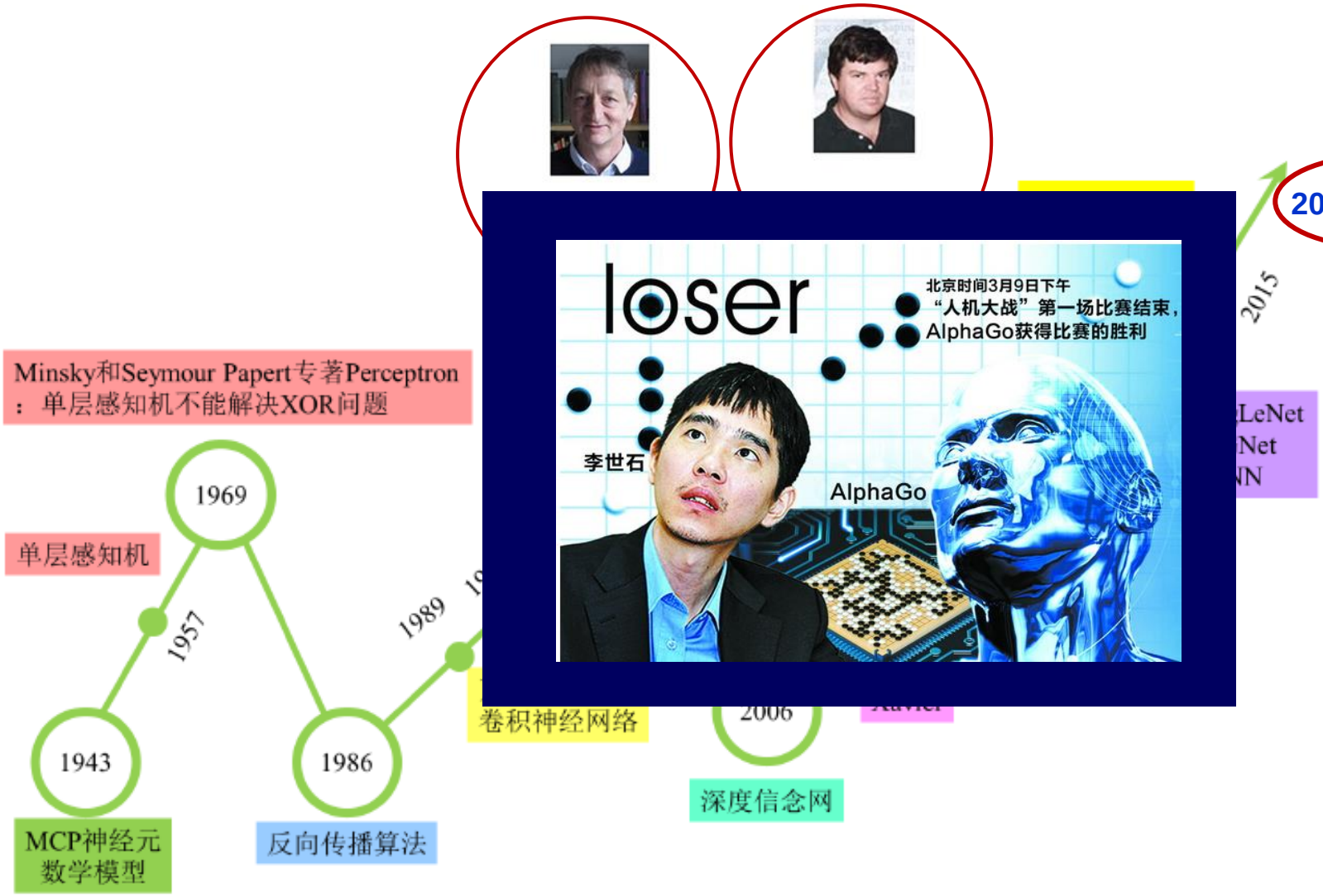
深度学习

2006年，著名的学者Geoffrey Hinton在Science上发表了一篇论文，给出了训练深层网络的新思路（无监督学习 、分层预训练、新的网络结构 、得名“深度”学习 ）

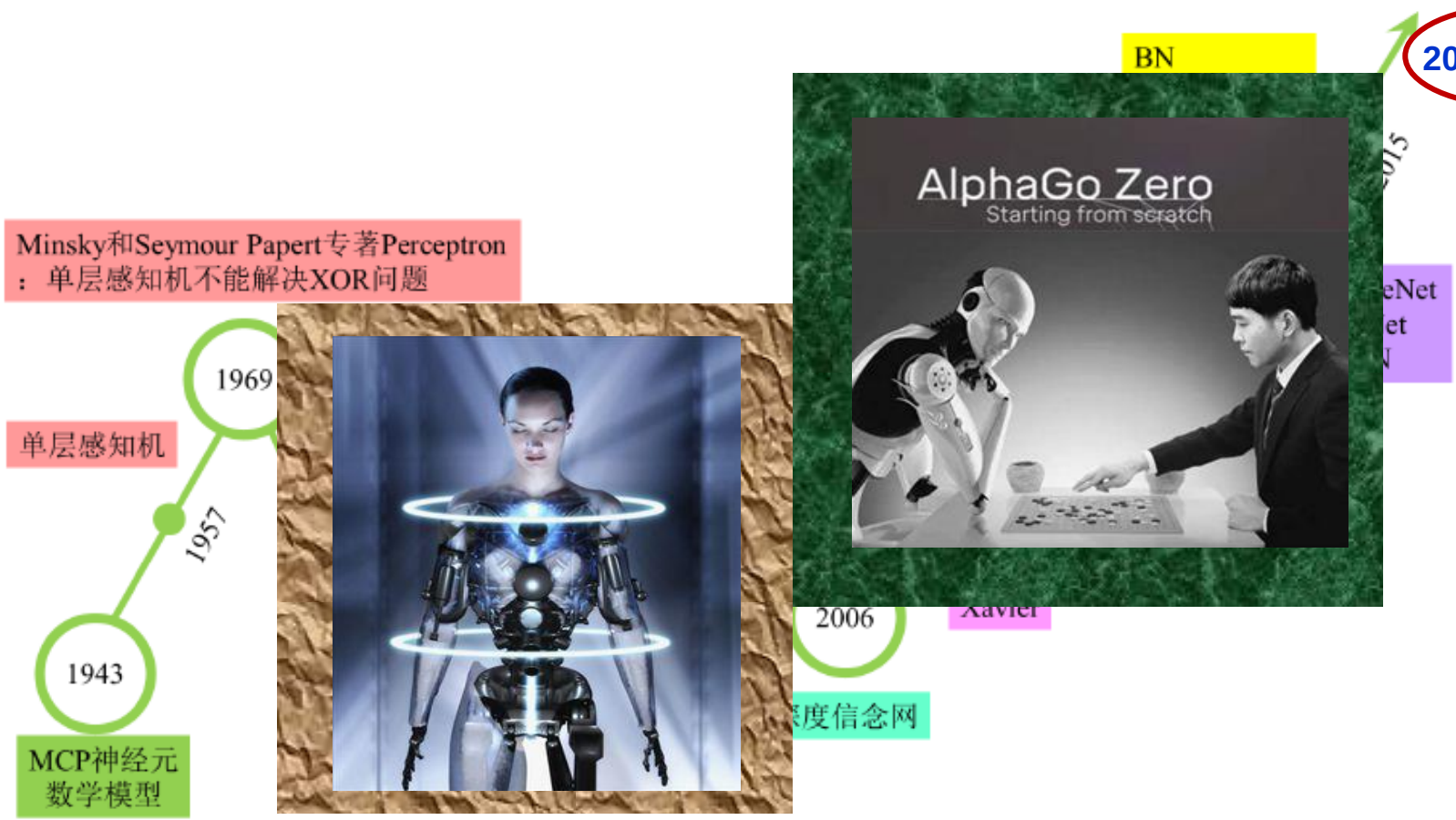
优化方法的突破是第三次NN研究浪潮兴起的钥匙

2012年底，Geoff Hinton的博士生Alex Krizhevsky、Ilya Sutskever（他们研究深度学习时间并不长）在图片分类的竞赛ImageNet上，识别结果拿了第一名。2011年冠军的准确率(top 5精度)是74.3%。2012年，Hinton和他的学生Alex等人参赛，把准确率一下提高到84.7%。靠着深度学习震惊了机器学习领域，从此大量的研究人员开始进入这个领域，一发不可收拾。截止到2015年5月份，ImageNet数据集的精度已经达到了95%以上，某种程度上跟人的分辨能力相当了。

深度学习的发展历史

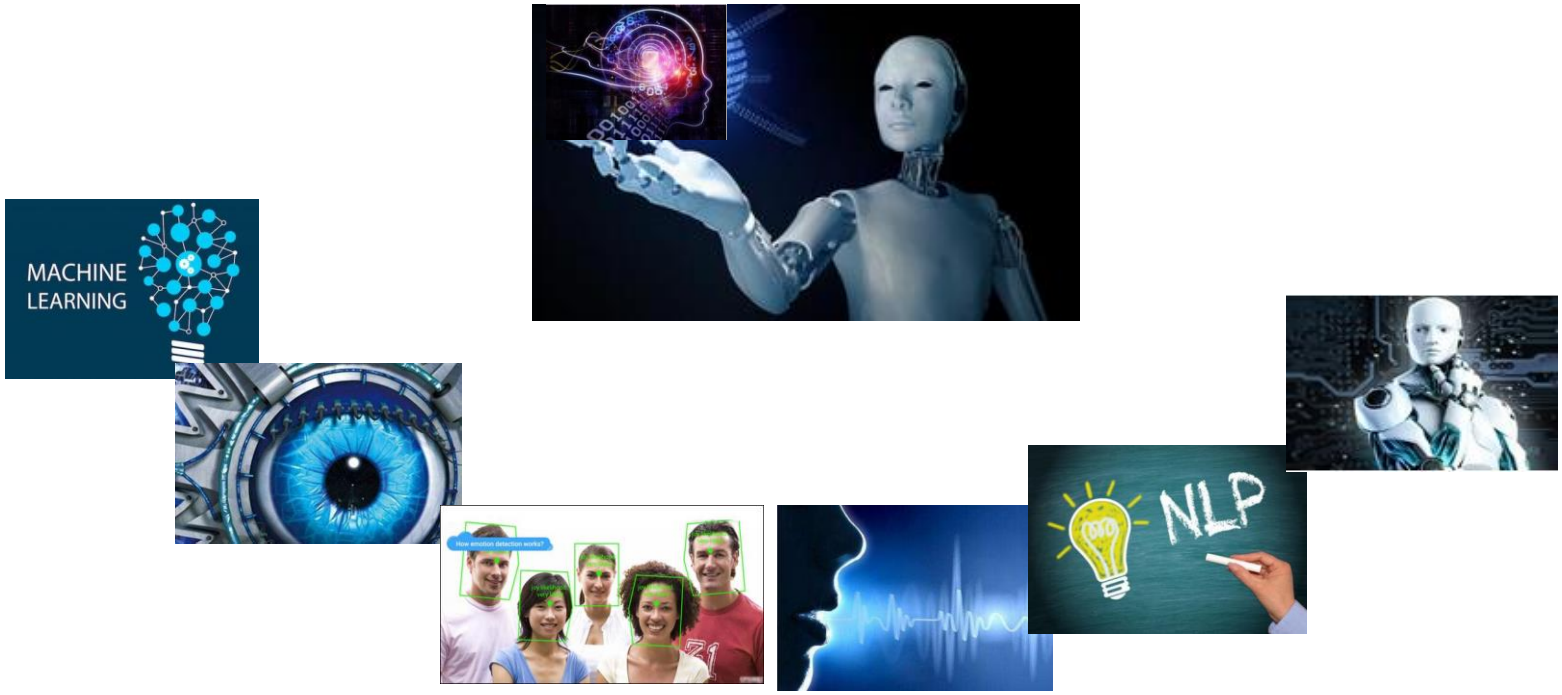

Minsky和Seymour Papert专著Perceptron：单层感知机不能解决XOR问题
1969
单层感知机
1957
1943
MCP神经元数学模型
1986
反向传播算法
1989
卷积神经网络
2006
深度信念网
2015
LeNet
Net
NN
loser
北京时间3月9日下午
“人机大战”第一场比赛结束，
AlphaGo获得比赛的胜利
李世石
AlphaGo
2016年3月

深度学习的发展历史


BN
2017年10月
AlphaGo Zero
Starting from scratch
Minsky和Seymour Papert专著Perceptron
：单层感知机不能解决XOR问题
1969
单层感知机
1957
1943
MCP神经元
数学模型
2006
eNet
et

深度学习的发展历史

深度学习（Deep Learning） 近年来火遍了各个领域



深度学习的发展历史

深度学习的重量级人物（2018年图灵奖）

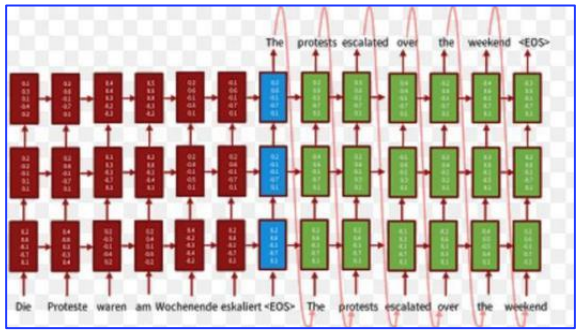
中： Geoff Hinton，deep learning 学派创始人

右： Yann Lecun，卷积神经网的发明者，Geoff Hinton的弟子

左： Yoshua Bengio，蒙特利尔大学

深度学习在NLP领域

近年来深度学习在NLP领域取得了许多重要成果



机器翻译——实现端到端的翻译模型，其优点是无复杂的中间环节设计，直接实现语言间的翻译。在30种语言上均比统计机器翻译模型的BLEU值有很大提高。

自动文摘——在深度学习的助力下，文本生成技术得到了进步，基于 RNN 模型在文本生成方面取得了很大的成就，也随之提升了抽象式文本摘要的效果

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize nigeria's economy. UNK says his administration is confident it will be able to thwart criminals and other nigerians. he says the country has long nigeria and nigeria's economy.

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption in the northeast part of nigeria. he says he'll "rapidly give attention" to curbing violence in the northeast part of nigeria. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

深度学习在NLP领域

智能问答——当前深度学习在解决问答领域中的关键问题取得了很大进展。代表产品：闲聊型个人助理：siri、微软小冰、微软cortana；智能客服：售前售后助理机器人（淘宝、京东）；

阅读理解——在深度学习的助力下，阅读理解技术得到了很大提升，机器阅读理解模型的评测结果甚至超过人类测评值。

2019年自然语言处理领域的大部分顶会中深度学习相关的论文占比达90%以上

深度学习在NLP领域

自然语言处理八个重大里程碑：

- ★ 2001年 - 神经语言模型
- ★ 2008年 - 多任务学习
- ★ 2013年 - Word嵌入
- ★ 2013年 - NLP的神经网络
- ★ 2014年 - 序列到序列模型
- ★ 2015年 - 注意力机制
- ★ 2015年 - 基于记忆的神经网络
- ★ 2018年 - 预训练语言模型

内 容 提 要

1. 自然语言处理与人工智能
2. 自然语言处理发展历史及学派
3. 自然语言处理技术及应用架构
4. 自然语言处理相关信息简介

3 自然语言处理技术及应用架构

✧ 自然语言处理架构

■ 自然语言处理任务

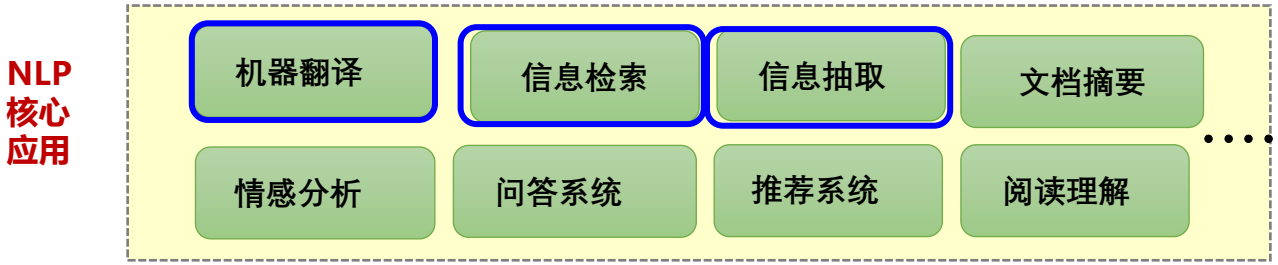
- NLP核心应用
- NLP+应用
- NLP基础技术

■ 自然语言处理方法及基础理论

- 规则法
- 概率统计法
- 深度学习法

3 自然语言处理技术及应用架构

■ 自然语言处理任务



机器翻译 (Machine translation, MT)：实现一种语言到另一种语言的自动翻译。

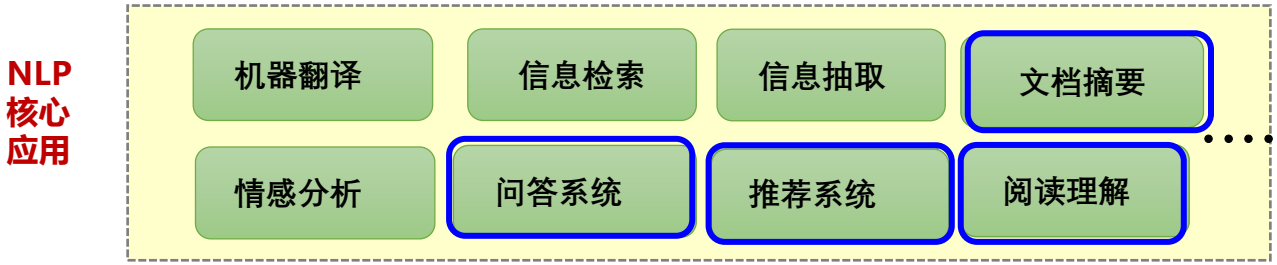
信息检索 (Information retrieval): 利用计算机系统从大量文档中找到用户需要的信息。

信息抽取 (Information extraction)：从指定文档中或者海量文本中抽取出用户感兴趣的信息。

文本分类 (Document classification)：利用计算机对大量的文档按照分类标准实现自动归类。

3 自然语言处理技术及应用架构

■ 自然语言处理任务



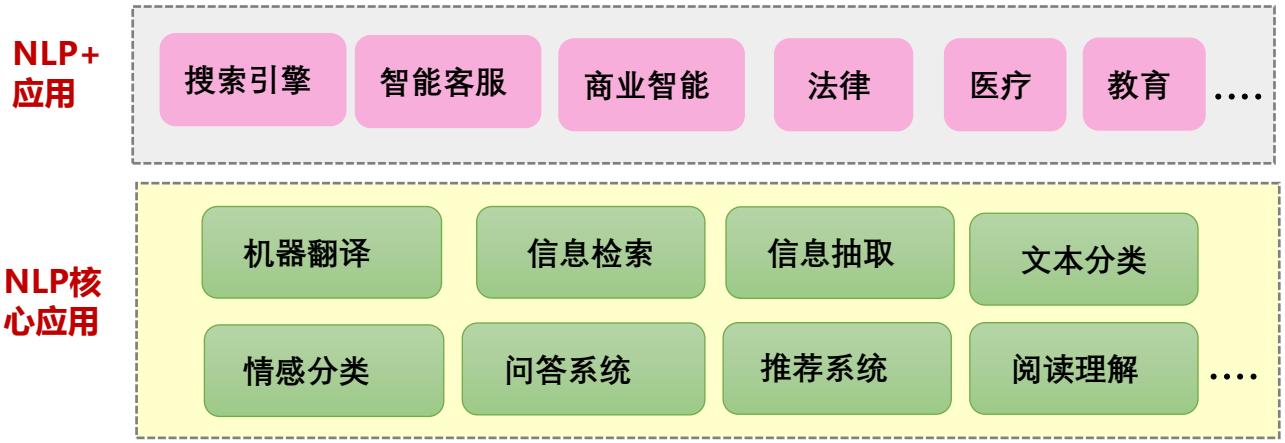
问答系统 (Question-answering system)：通过计算机系统对人提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并做出相应的回答。

推荐系统(Recommendation system)：根据用户的习惯、偏好或兴趣，从大规模信息中识别满足用户兴趣的信息并推荐给用户。

阅读理解 (Machine Reading)：要求系统回答一些非事实性的、高度抽象的问题。同信息源被限定于给定的一篇文章，相对于传统问答任务，机器阅读理解更具挑战。

1. 自然语言处理与人工智能

自然语言处理任务



NLP+应用：NLP技术与领域深度结合，将自然语言处理技术深入到各个应用系统和垂直领域中，为行业创造价值。如，智能客服、智能教育、法律助手、智能医疗系统、商业智能……

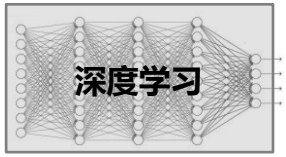
3 自然语言处理技术及应用架构

理性主义： 1960s – 1980s中期

- 以语言学家 N.Chomsky 为代表。认为人类生成合乎文法的语句的能力是生来具有的,为此他提出一种称为生成句法(Generative Grammar) 的理论
- 理性主义试图去描写人脑中的语言模型（产生模型）：通过一组有限的规则作用于一个有限的词汇集, 生成无限的可接受的、合乎文法的句子
- 技术路线：人工规则方法
- 理论基础：基于乔姆斯基的语言理论

- 以行为心理学家伯尔赫斯·弗雷德里克·斯金纳B. F. Skinner为代表。认为人类语言能力的获得来自于学习,语言是通过不断地实践而“约定俗成”的结果。
- 经验主义试图去刻画真实世界的语言现象
- 技术路线：统计学习方法（数据驱动-从数据中学习的方法）
- 理论基础：基于香农的信息论--将语言事件赋予概率，作为其可信度，由此来判断一个句子是常见的还是罕见的
- 要素：数学基础、统计算法、训练语料

1920s~1950s/1980s中期



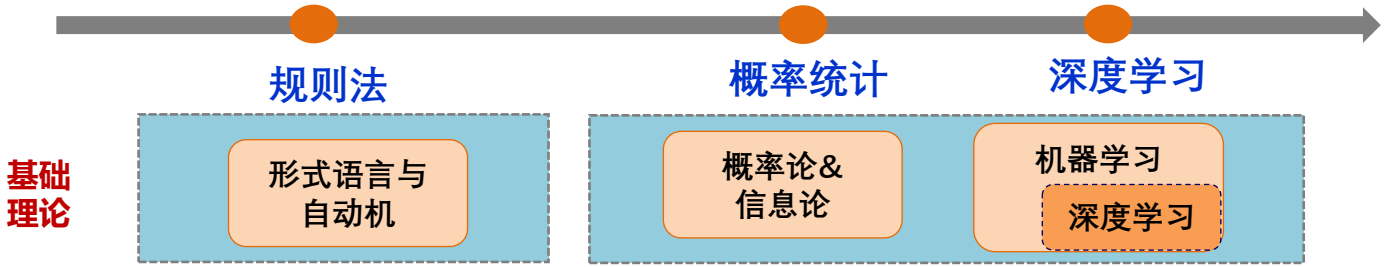
2010s~至今

- 仍属于经验主义学派
- 技术路线：神经网络方法（数据驱动-从数据中学习的方法）
- 理论基础：深度学习理论及方法
- 要素：数学基础、深度学习算法、训练语料
- 特点：端到端的解决问题方式

经验主义：

3 自然语言处理技术及应用架构

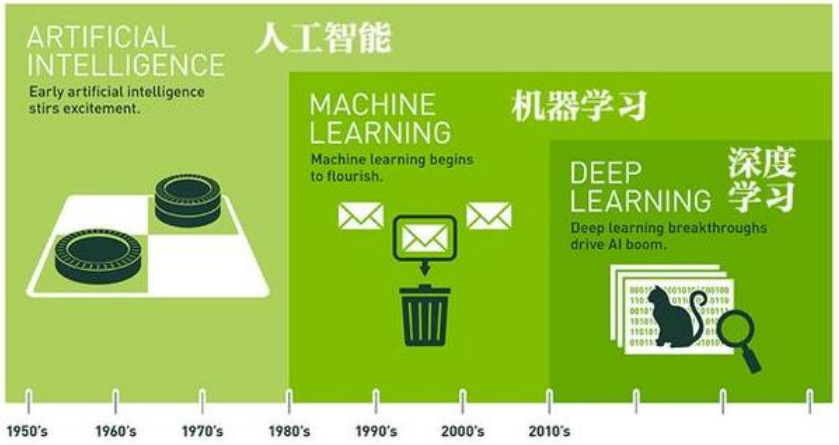
■ 自然语言处理方法及基础理论



- ❖ **形式语言与自动机**
  - ➢ Chomsky形式语言

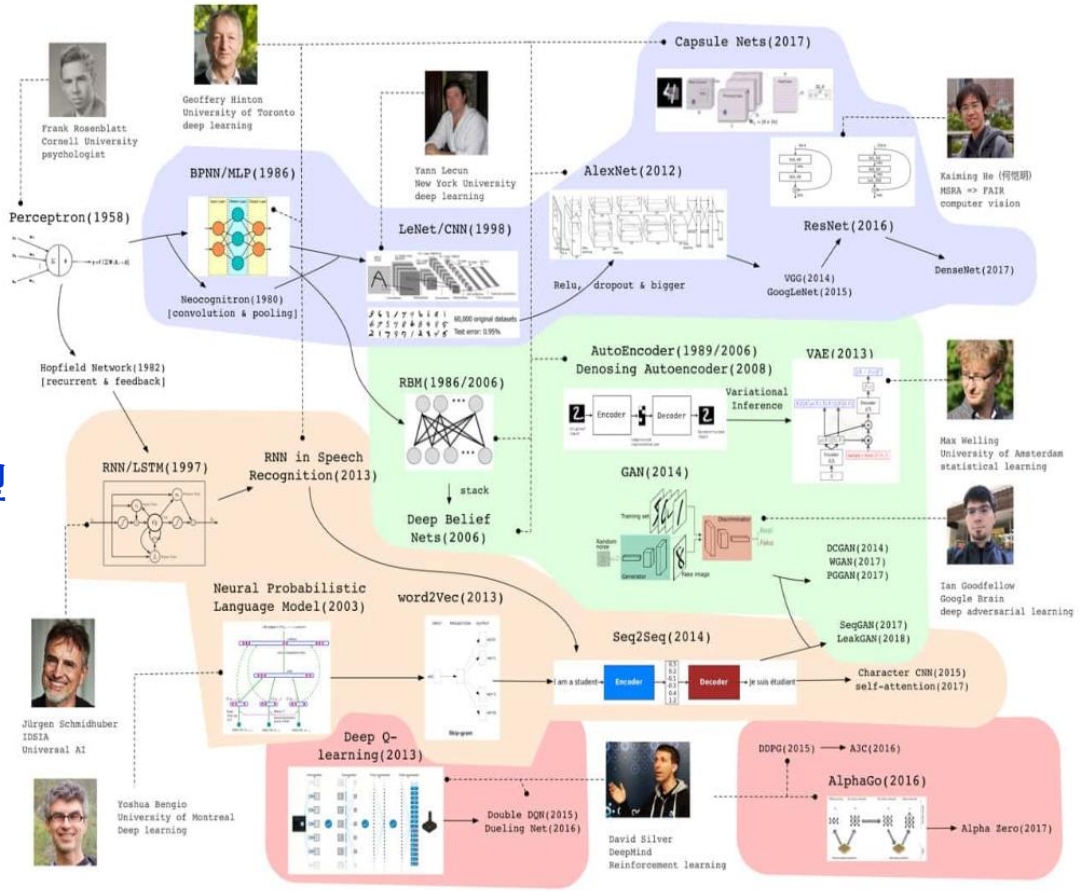
- ❖ **概率论&信息论**
  - ➢ 概率论
  - ➢ 信息论

- ❖ **机器学习：**
  - ➢ 概率图模型
  - ➢ 神经网络



3 自然语言处理技术及应用架构

深度学习的四个主要脉络



卷积网络

无监督学习

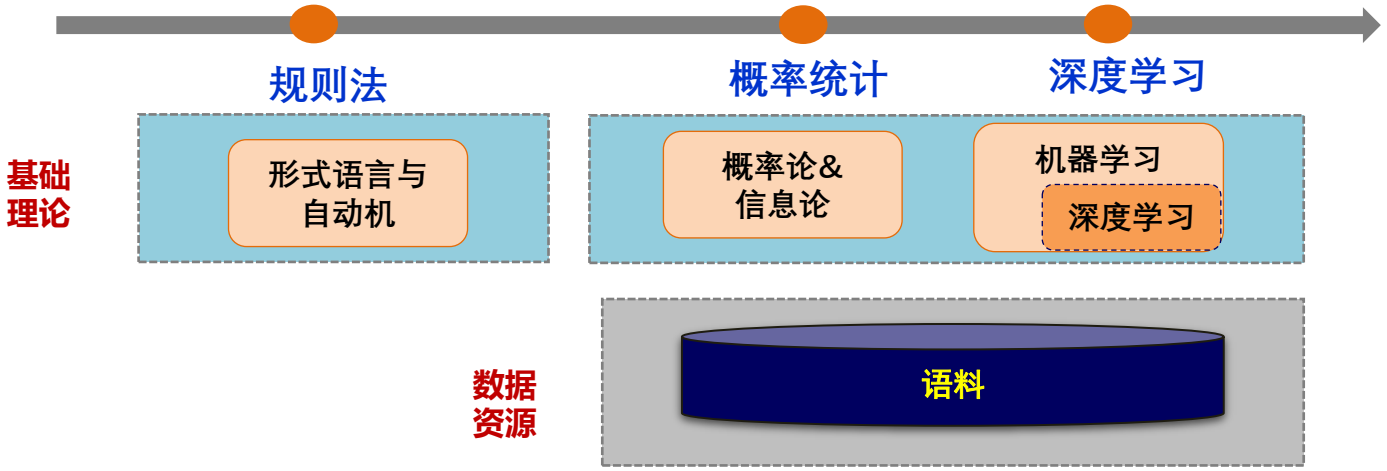
序列深度模型

增强学习

深度学习模型最近若干年的重要进展

3 自然语言处理技术及应用架构

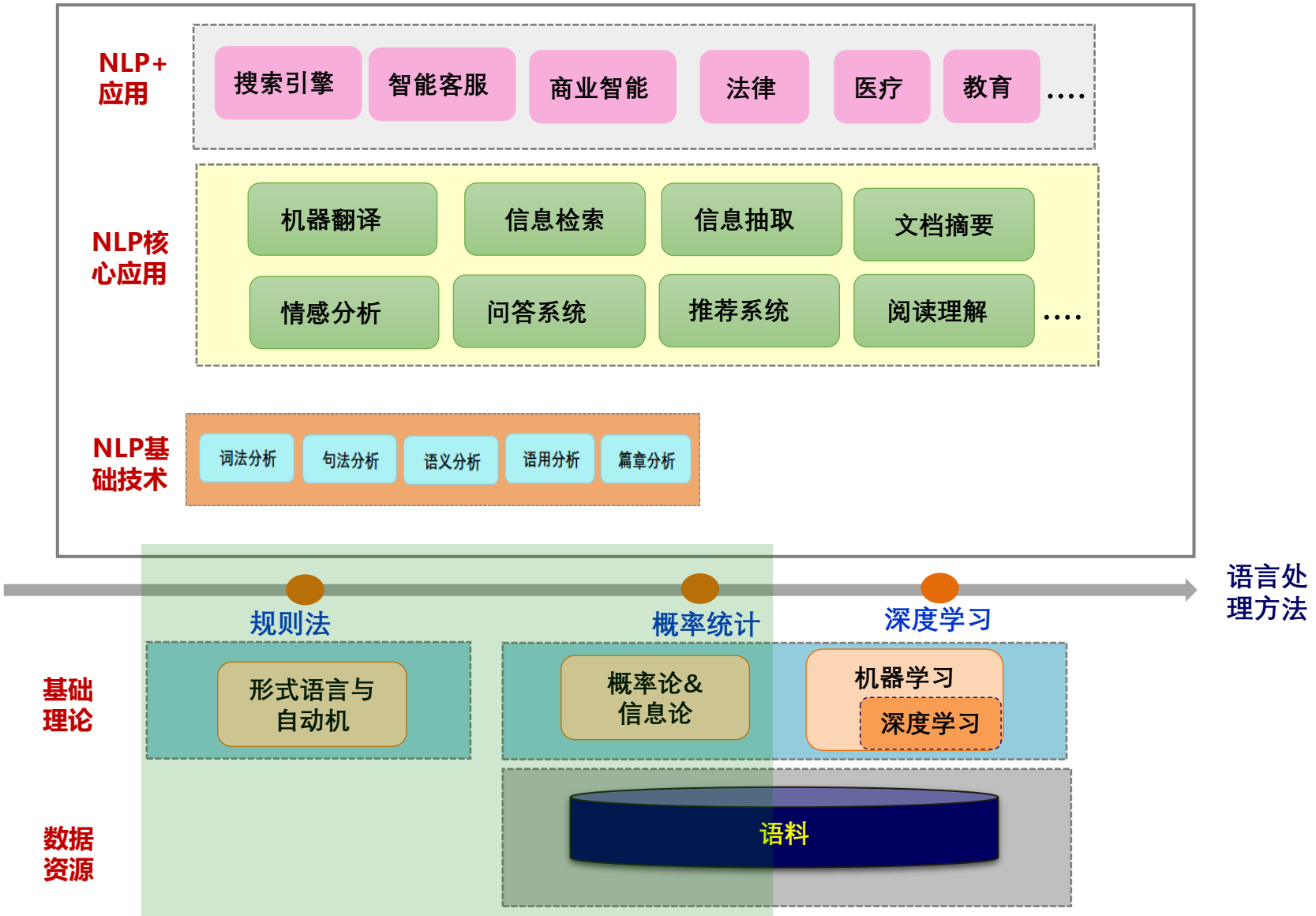
■ 自然语言处理方法及基础理论



- **传统语料库（corpus base）：** 按照一定的原则组织在一起的真实的自然语言数据(包括书面语和口语)的集合是统计NLP的知识来源
- **语言知识库：** 按照一定的原则组织在一起的人类加工处理后的语言知识库
- **任务数据集：** 按照任务建立的标注（未标注）的数据集

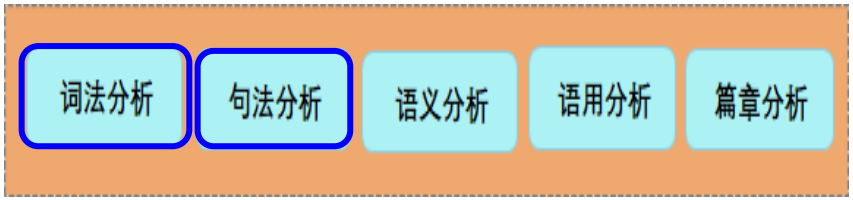
3 自然语言处理技术及应用架构

传统的自然语言处理课程体系



3 自然语言处理技术及应用架构

NLP基础技术



词法分析：从句子中分出单词，从中获得单词的语言学信息并确定单词的词性。

词法分析是很多中文信息处理任务的必要步骤。

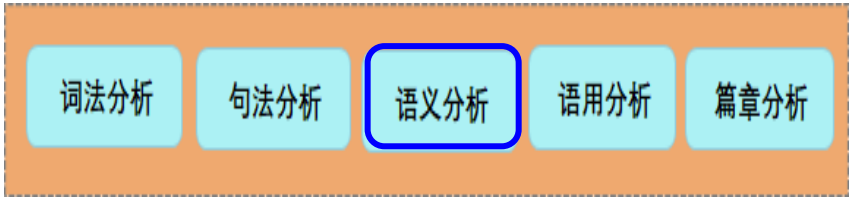
- 具体包括：
  - 自动分词（中文分词）
  - 命名实体识别
  - 词性标注

句法分析： 句法分析是对句子和短语结构进行分析，如句子的形式结构：主语、谓语、宾语等。句法分析是语言学理论和实际的自然语言应用的一个重要桥梁。

- **短语结构分析** （主要采用宾州树库）
- **依存句法分析** （图的分析算法和基于转换的分析算法）

3 自然语言处理技术及应用架构

NLP基础技术



语义分析： 解释自然语言句子或篇章各部分(词、词组、句子、段落、篇章)的意义。目前语义计算的理论、方法、模型尚不成熟

- **词义排歧（词层次）**

  确定一个多意词在给定的上下文语境中的具体含义

- **语义角色标注 （句子层次）**

  为句子中的每个动词标注出其相关的名词及其语义角色（属于浅层语义分析技术）

3 自然语言处理技术及应用架构

NLP基础技术

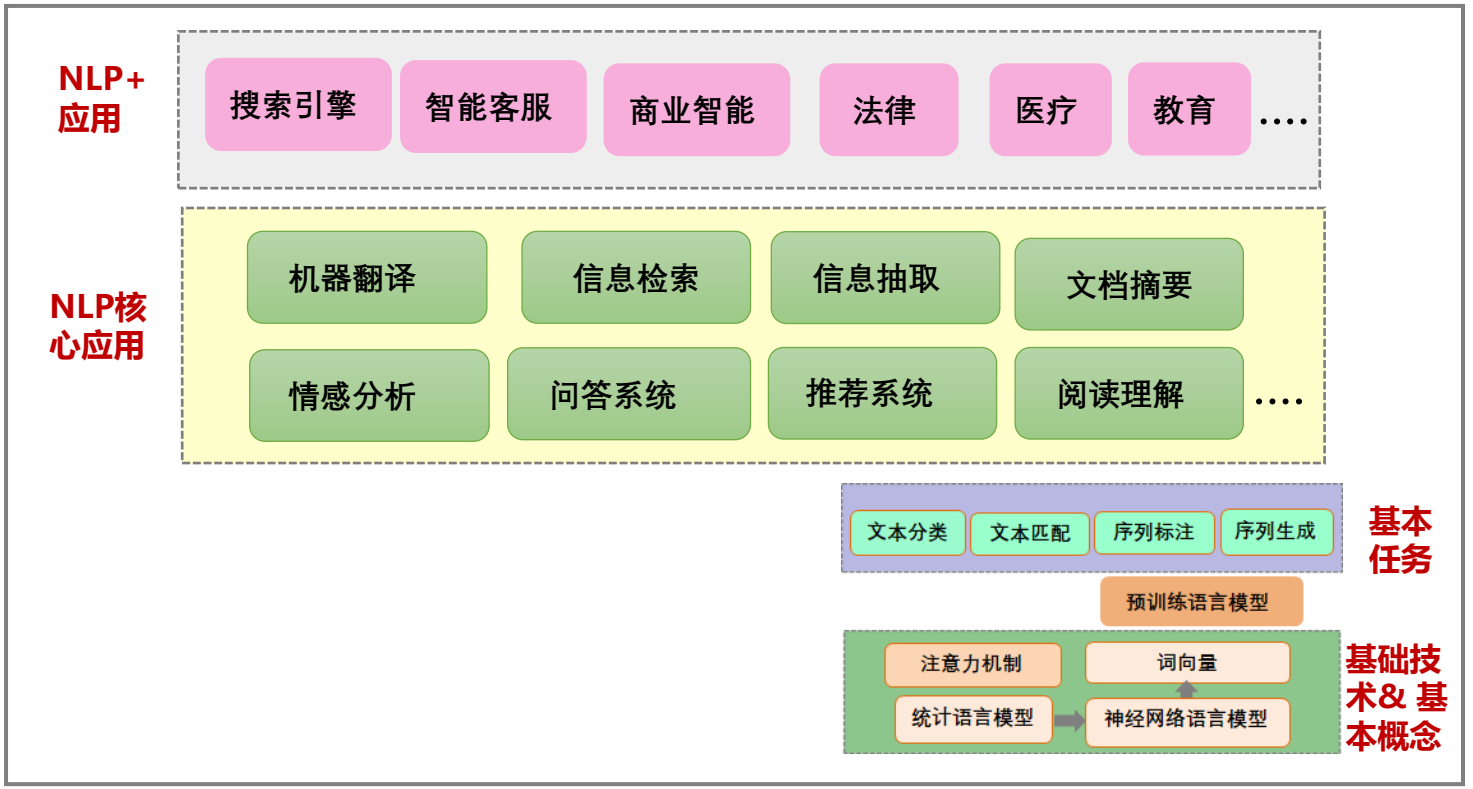
语用分析： 研究语言所在的外界环境对语言使用所产生的影即说话双方按照该单词或者语言成分所在的“语境”，来确定应该选择其中哪一种释义或含义。目前语用的理论、方法、模型尚不成熟

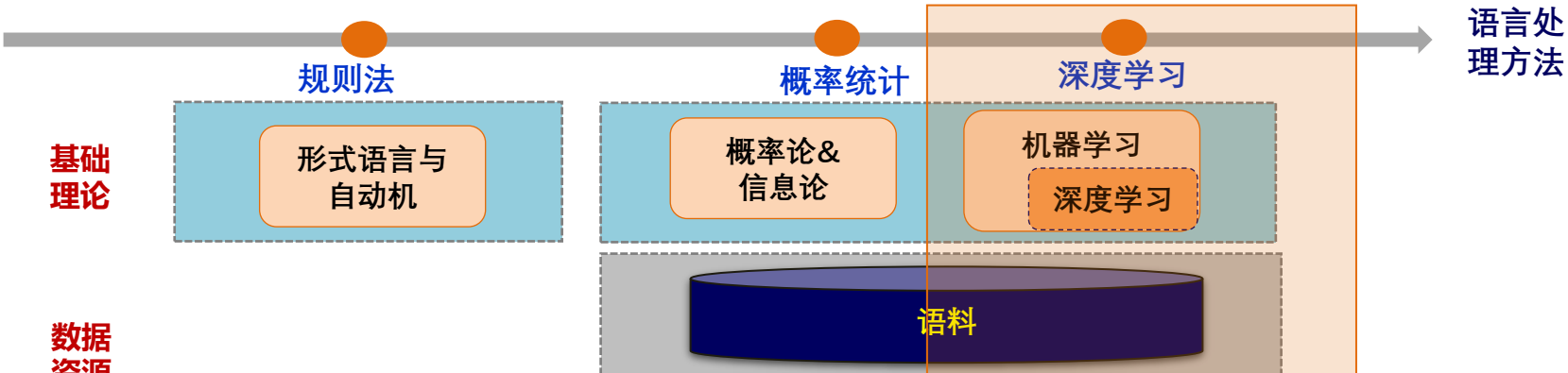
篇章分析： 句子（语段）之间的关系以及关系类型的划分，段落之间的关系的判断，跨越单个句子的词与词之间的关系分析，话题的继承与变迁等。

- 篇章连贯性分析
- 篇章衔接性分析

3 自然语言处理技术及应用架构

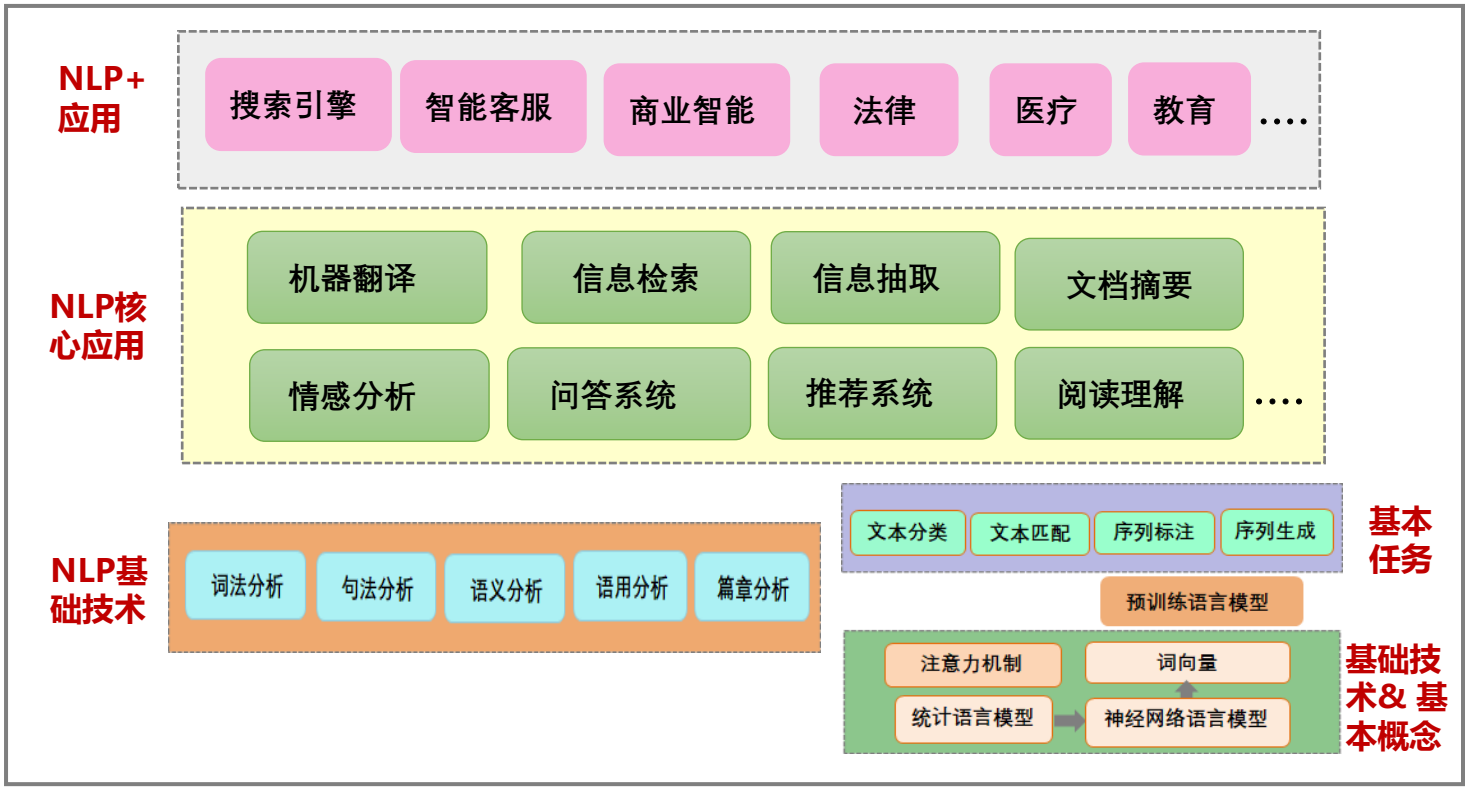
基于深度学习的自然语言处理课程体系

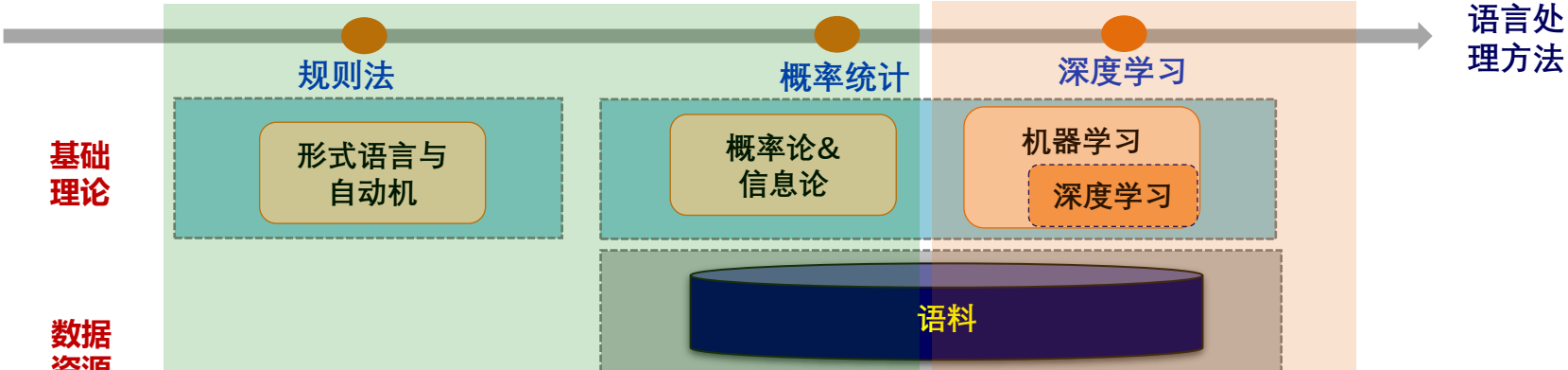




3 自然语言处理技术及应用架构

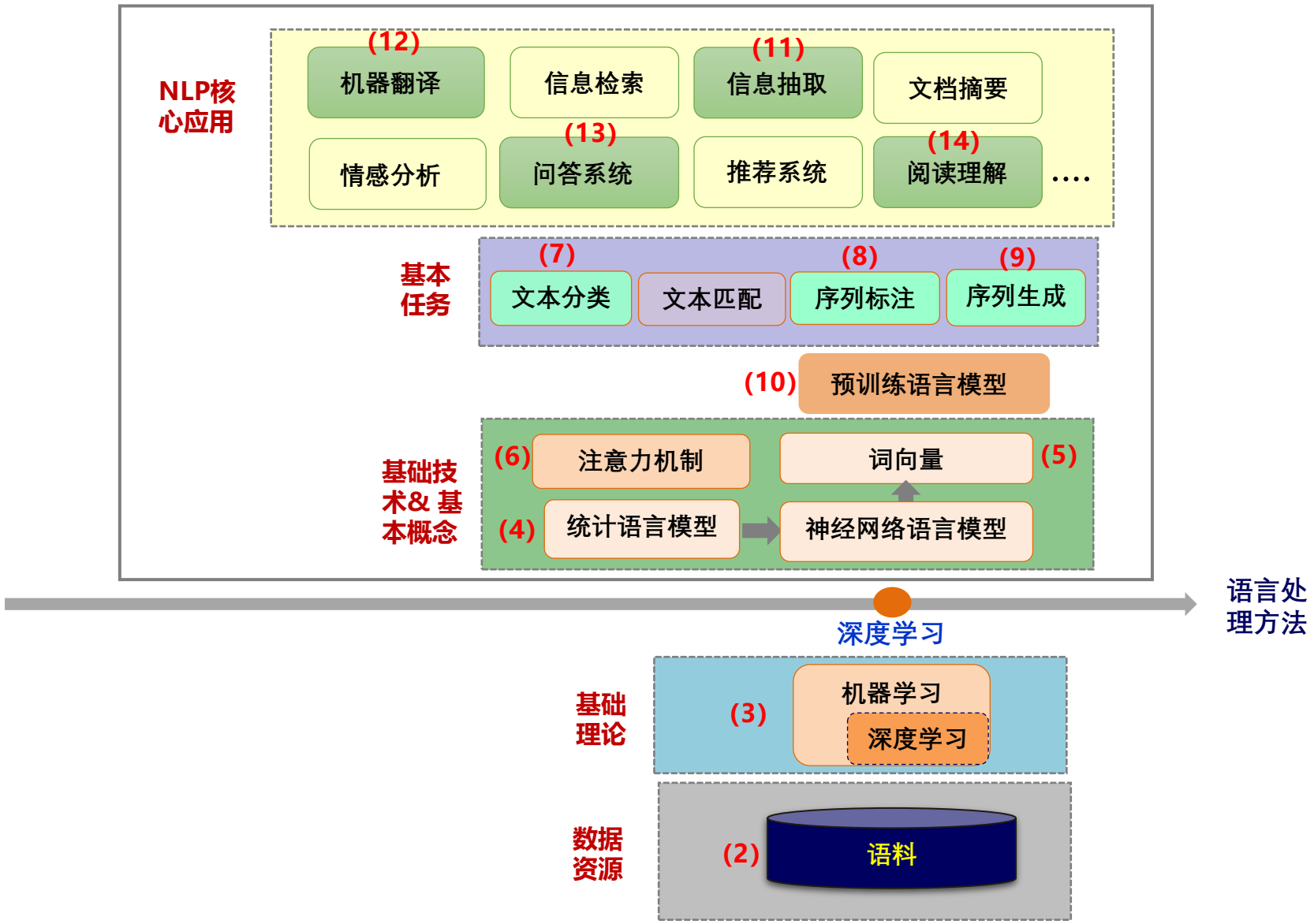
自然语言处理体系架构





3 自然语言处理技术及应用架构

基于深度学习的自然语言处理授课体系



内 容 提 要

1. 自然语言处理与人工智能
2. 自然语言处理发展历史及学派
3. 自然语言处理技术及应用架构
4. 自然语言处理相关信息简介

4. 自然语言处理技术评测及学术会议

□ NLP领域评测

❖ NIST系列评测(National Institute of Standard and Technology)

在美国国防先进技术研究计划署（DARPA，Defense Advanced Research Projects Agency ）等部门支持下，开展了一系列周期性的技术评测工作，**目前为止国际上影响力最大的系列评测。**

- 语音识别系列评测
- 文本检索评测TREC
- 机器翻译评测（Open MT Evaluation）
- 信息提取评测（MUC、ACE）
- 话题检测与跟踪评测（TDT ）
- 多文档文摘评测（DUC）

4. 自然语言处理技术评测及学术会议

- **句法分析其他国际评测：**
  - 中文分词：SIGHAN Chinese Language Processing Bakeoff
  - 跨语言检索：NTCIR，CLEF
  - 机器翻译：IWSLT，TCSTAR
  - 语言分析：CoNLLShared Task
  - 语义处理：SemEval

- **不同机构组织的各种竞赛**
  - 机器翻译
  - 知识图谱
  - 阅读理解
  - 情感分析
  - .....

4. 自然语言处理技术评测及学术会议

□ NLP领域的学术会议

主要国际会议

– ACL（Association of Computational Linguistics）

– EMNLP （ Conference on Empirical Methods in Natural language Processing）

– NAACL （The North American Chapter of the Association for Computational Linguistics）

– Coling （ International Conference on Computational Linguistics）

– EACL(European Chapter of ACL)

– IJCNLP(International Joint Conference on Natural language Processing)

– SIGIR(SIG Information Retrieval)

– TREC(Text REtrieval Conference)

……

主要国内会议

– JSCL(全国计算语言学联合学术会议)

4. 自然语言处理技术评测及学术会议

□ 国内外自然语言处理(NLP)研究组

中国大陆地区：

腾讯人工智能实验室（Tencent AI Lab）
https://ai.tencent.com/ailab/nlp/
苏州大学自然语言处理实验室
http://nlp.suda.edu.cn/
微软亚洲研究院自然语言计算组 Natural Language Computing (NLC) Group
https://www.microsoft.com/en-us/research/group/natural-language-computing/
头条人工智能实验室（Toutiao AI Lab）
http://lab.toutiao.com/
清华大学自然语言处理与社会人文计算实验室
http://nlp.csai.tsinghua.edu.cn/site2/
清华大学智能技术与系统国家重点实验室信息检索组
http://www.thuir.cn/cms/
北京大学计算语言学教育部重点实验室
http://www.klcl.pku.edu.cn/
北京大学计算机科学技术研究所语言计算与互联网挖掘研究室
http://www.icst.pku.edu.cn/lcwm/index.php?title=%E9%A6%96%E9%A1%B5
哈工大社会计算与信息检索研究中心
http://ir.hit.edu.cn/

4. 自然语言处理技术评测及学术会议

哈工大机器智能与翻译研究室

http://mitlab.hit.edu.cn/

哈尔滨工业大学智能技术与自然语言处理实验室

http://www.insun.hit.edu.cn/home/

中科院计算所自然语言处理研究组

http://nlp.ict.ac.cn/index_zh.php

中科院自动化研究所语音语言技术研究组

http://nlpr-web.ia.ac.cn/cip/introduction.htm

南京大学自然语言处理研究组

http://nlp.nju.edu.cn/homepage/

复旦大学自然语言处理研究组

http://nlp.fudan.edu.cn/

东北大学自然语言处理实验室

http://www.nlplab.com/

厦门大学智能科学与技术系自然语言处理实验室

http://nlp.xmu.edu.cn/

4. 自然语言处理技术评测及学术会议

北美:

Natural Language Processing - Research at Google
https://research.google.com/pubs/NaturalLanguageProcessing.html
Facebook AI Research (FAIR)
https://research.fb.com
Thomas J. Watson Research Center -IBMResearch
http://researchweb.watson.ibm.com/labs/watson/index.shtml
The Stanford Natural Language Processing Group
http://nlp.stanford.edu/
The Berkeley NLP Group
http://nlp.cs.berkeley.edu/index.shtml
Artificial Intelligence Research Group at Harvard
http://www.eecs.harvard.edu/ai/
The Harvard natural-language processing group
http://nlp.seas.harvard.edu/
Natural Language Processing Group at MIT CSAIL
http://nlp.csail.mit.edu/
Human Language Technology Research Institute at University of Texas at Dallas
http://www.hlt.utdallas.edu/

4. 自然语言处理技术评测及学术会议

Natural Language Processing Group at Texas A&M University
http://nlp.cs.tamu.edu/
The Natural Language Processing Group at Northeastern University
https://nlp.ccis.northeastern.edu/
Cornell NLP group
https://confluence.cornell.edu/display/NLP/Home/
Natural Language Processing group at University Of Washington
https://www.cs.washington.edu/research/nlp
Natural Language Processing Research Group at University of Utah
https://www.cs.utah.edu/nlp/
Natural Language Processing and Information Retrieval group at University of Pittsburgh
http://www.isp.pitt.edu/research/nlp-info-retrieval-group
Brown Laboratory for Linguistic Information Processing (BLLIP)
http://bllip.cs.brown.edu/

4. 自然语言处理技术评测及学术会议

- **NLP课程资源：**

学校	课程名	网址
CMU	Natural Language Processing	http://demo.clab.cs.cmu.edu/NLP/
MIT	Natural Language Processing	http://web.mit.edu/6.863/www/fall2012/
Stanford University	Natural Language Processing	http://online.stanford.edu/course/natural-language-processing
Columbia University	Natural Language Processing	http://www.cs.columbia.edu/~cs4705/

4. 自然语言处理技术评测及学术会议

NLP 推荐信息：

- 公众号：
  - 机器之心、专知、新纪元、 Paperweekly、哈工大SCIR、学术头条
- 博客：
  - Sebastian Ruder: http://ruder.io/#open
  - 知乎专栏：https://zhuanlan.zhihu.com/xitucheng10

参考文献

冯志伟，自然语言处理的历史与现状

宗成庆，统计自然语言处理（第2版）

刘昕 ， 深度学习一线实战暑期研讨班 ----深度学习基础

周明，自然语言处理的历史与未来

人工智能影响力报告： http://www.sohu.com/a/132070214_505794

中文信息处理发展报告，中国中文信息学会，2016

洞察一文带你全面了解自然语言处理发展史上的8大里程碑
http://dy.163.com/v2/article/detail/DV2I22QO051480KF.html

在 此 表 示 感 谢!

谢谢各位！

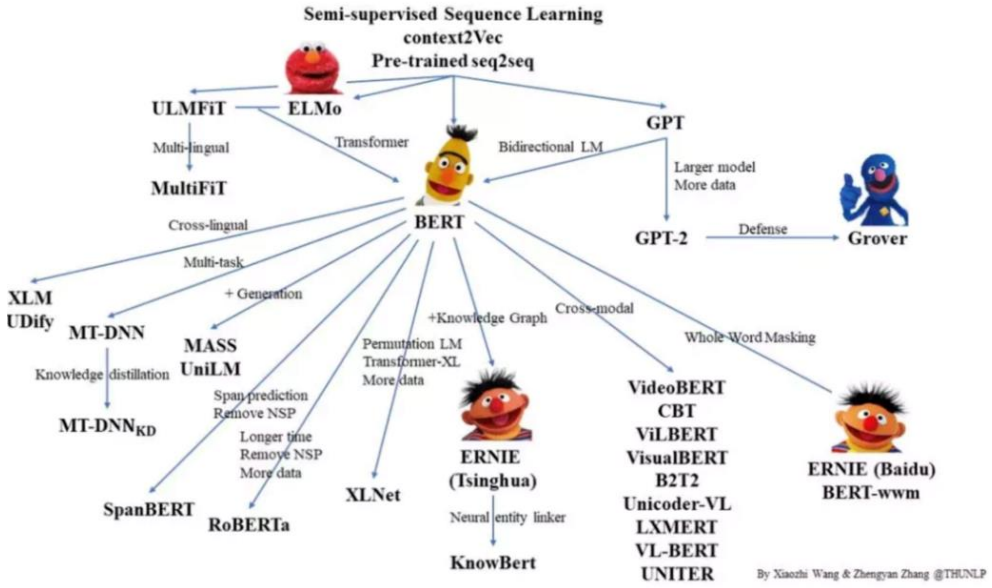
Q&A

附录：当前自然语言处理热点研究问题

据 ACL-IJCAI-SIGIR （AIS 2020）顶级会议论文报告数据显示

当前自然语言处理热点研究问题：

■ 预训练语言模型



1.无监督预训练

2.跨语言学习

当前自然语言处理热点研究问题：

■ 问答

将知识图谱和QA相结合，以实现更复杂的QA；QA中的推理和多跳推理问题，QA通常还会和其它任务结合，形成多任务框架，以提升多个任务的效果

当前自然语言处理热点研究问题：

■ 自然语言生成

自然语言生成有着广阔的应用前景，也是近年来的研究热点

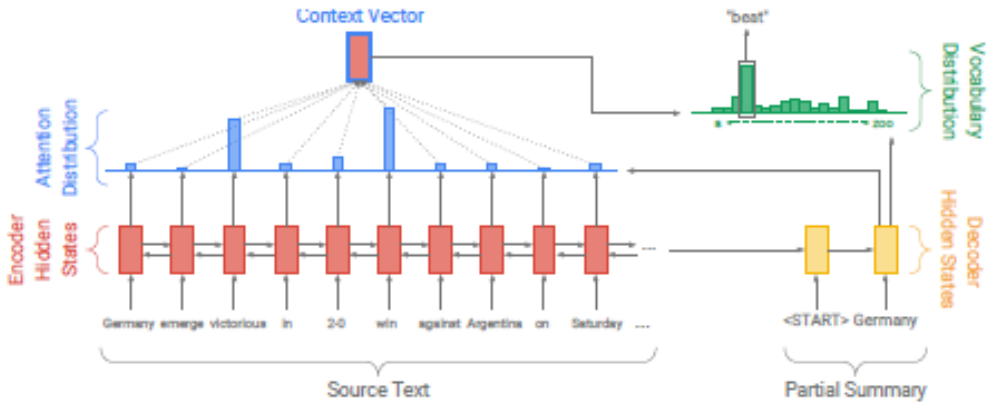
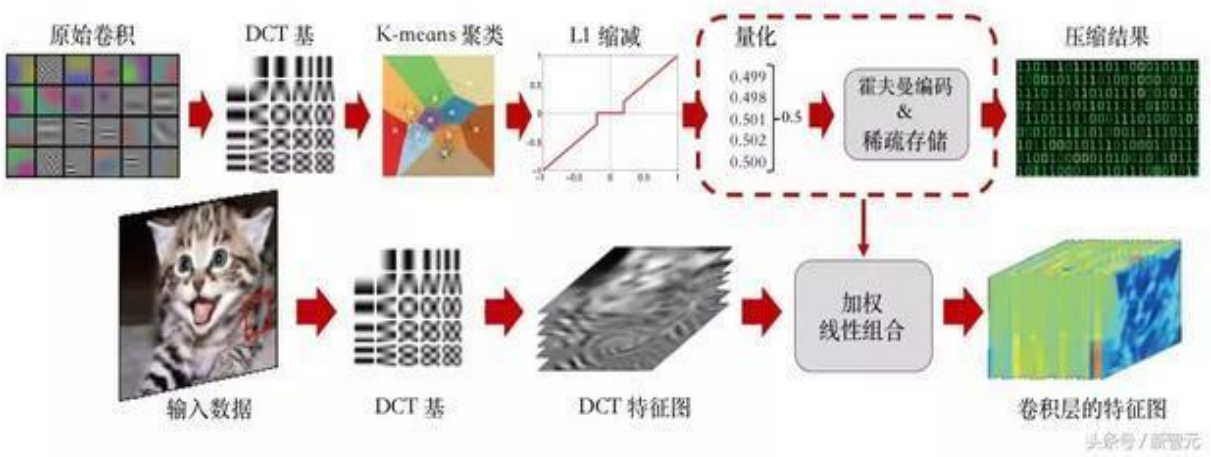


Figure 2: Baseline sequence-to-sequence model with attention. The model may attend to relevant words in the source text to generate novel words, e.g., to produce the novel word *beat* in the abstractive summary *Germany* ***beat*** *Argentina 2-0* the model may attend to the words *victorious* and *win* in the source text.

当前自然语言处理热点研究问题：

■ 多模态

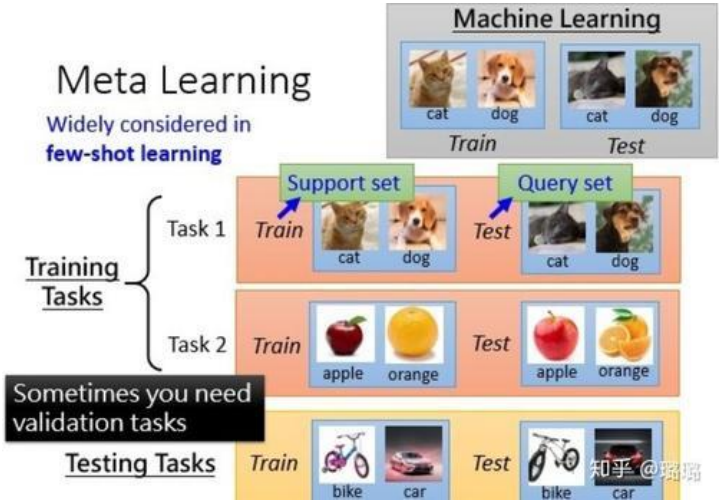
将文本和其它的语音、视频、图像的模态相结合

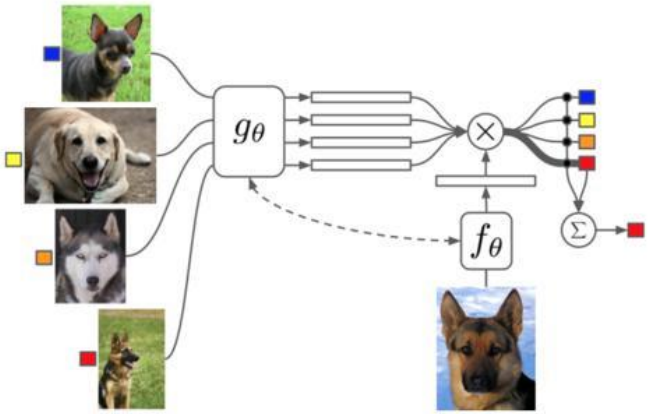


当前自然语言处理热点研究问题：

自然语言处理热点研究问题（算法层面）：

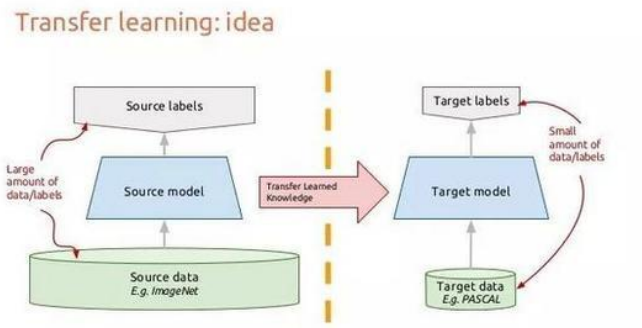
- 元学习和少样本学习



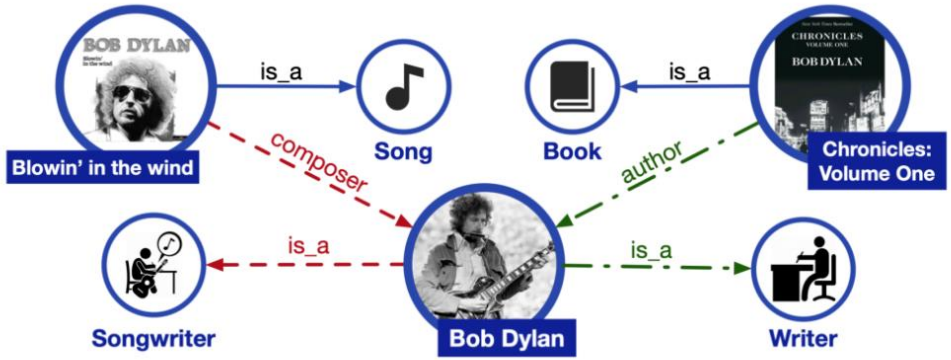


当前自然语言处理热点研究问题：

- 迁移学习



- 知识融合



当前自然语言处理热点研究问题：

■ 误差

在NLP领域中，由于数据集不均衡的原因，以及各种各样的固有偏见，会出现各种各样的Bias，如何消除这种bias对NLP算法来说至关重要