自然语言处理

中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

2018.09

课程信息：

主讲教师

胡玥，中科院信工所，研究员，博导

讲授内容：自然语言基础部分，机器翻译，阅读理解

邮箱：huyue@iie.ac.cn

王厚峰，北京大学，教授，博导

讲授内容：信息抽取，问答系统

邮箱：wanghf@pku.edu.cn

络卫华，阿里达摩研究院，资深算法专家，博导

讲授内容：NLP前沿技术

邮箱：weihua.luowh@alibaba-inc.com

助教老师

于静，中科院信工所，助理研究员

邮箱：yujing02@iie.ac.cn

课程信息：

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

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：

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

参考教材：

统计自然语言处理（第2版） 宗成庆， 清华大学出版社

Neural Network Methods for Natural Language Processing ， Yoav Goldberg, MORGAN &CLAYPOOL PUBLISHERS

参考资料

- 统计自然语言处理基础 ，[美] Chris Manning 电子工业出版社
- 自然语言理解（第二版），[美] James Allen,电子工业出版社
- 自然语言处理综论，[美]·Daniel Jurafsky,电子工业出版社

课程信息：

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

课程安排

上课时间：周二下午1:30-4:20

日期：2018.9.11 - 2019.1.15

总学时： 60学时/3学分

考核方式

平时作业：50%

结业考试：50%

第 1 章　绪 论

中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

内 容 提 要

1. 自然语言处理概述

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. 自然语言处理技术评测及学术会议

内 容 提 要

1. 自然语言处理概述

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. 自然语言处理技术评测及学术会议

1. 自然语言处理概述

✧ 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. 自然语言处理概述

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

- 本课程主要研究内容

书面形式的汉语文本

1. 自然语言处理概述

✧ 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

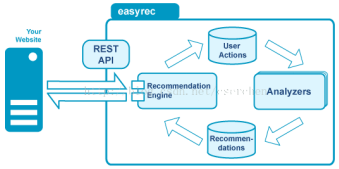
其它名称

- **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. 自然语言处理概述

✧ 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

1. 自然语言处理概述

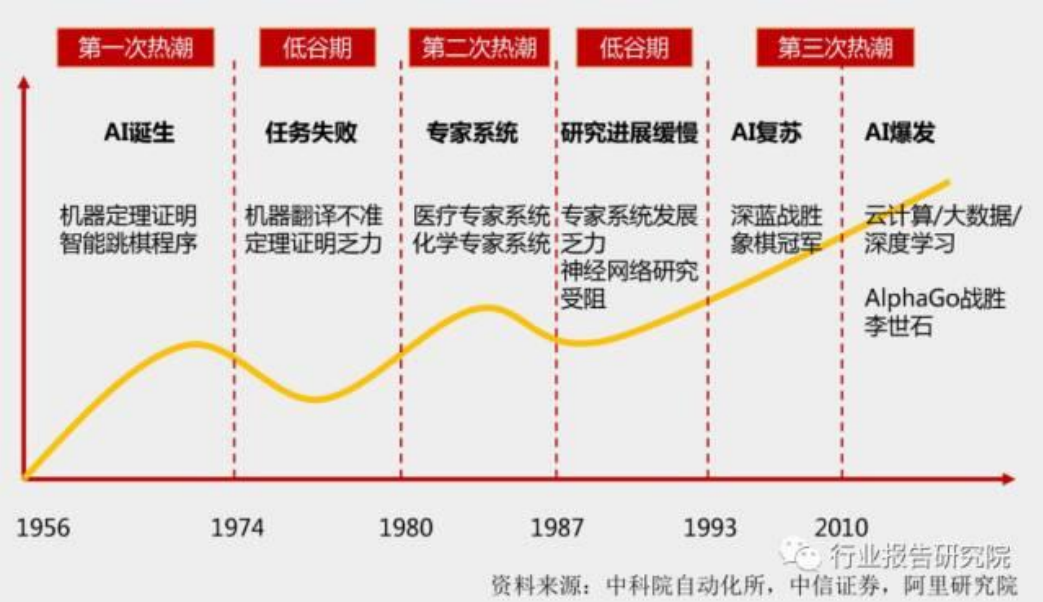
✧ 自然语言处理大背景

人工智能“新浪潮”大背景给自然语言处理带来新的机遇和挑战

人工智能（AI）— 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。

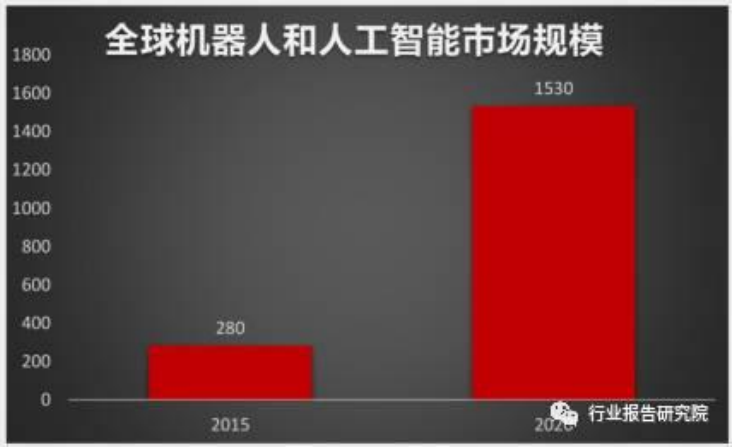
人工智能发展的三次热潮



1. 自然语言处理概述

人工智能市场未来5年呈现井喷趋势

根据美银美林估计，2020年，全球机器人和人工智能市场规模将达1530亿美元。



人工智能将在以下领域产生深远的影响

1. 自然语言处理概述

✧ 人工智能产品

智能客服机器人

阿里小蜜

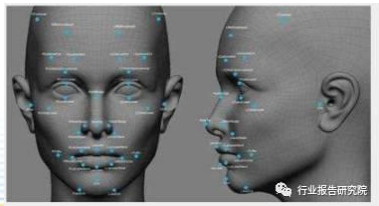
AI+零售

AI +交通

AI+金融

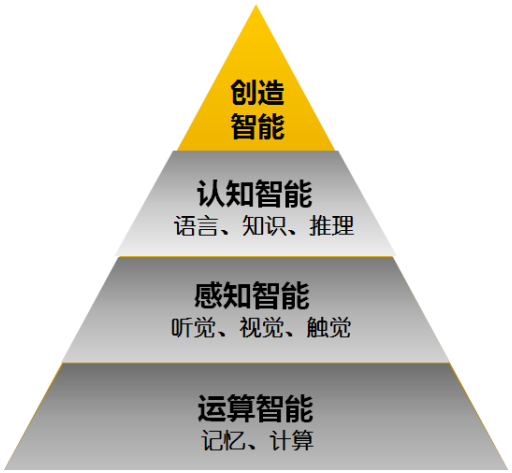
AI+制造

计算机视觉让机器看懂人和物



1. 自然语言处理概述

✧ 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；

最高一层是创造智能：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. 自然语言处理概述

✧ 人工智能关键技术

现状：最近两年，随着深度学习的引入，大幅度提高语音识别和图像识别的识别率，在一些典型的测试题下，达到或者超过了人类的平均水平；**但自然语言处理的水平**没有达到这种高度，还有许多问题亟待解决。

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

内 容 提 要

1. 自然语言处理概述
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. 自然语言处理技术评测及学术会议

2. 自然语言处理发展历史及学派

理性主义：1960s – 1980s中期

优点：
- 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

缺点：
- 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

优点：
- 统计模型提供了统一的冲突解决机制
- 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

缺点：
- 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）没有好的解决办法

1920s - 1950s，1980s中期

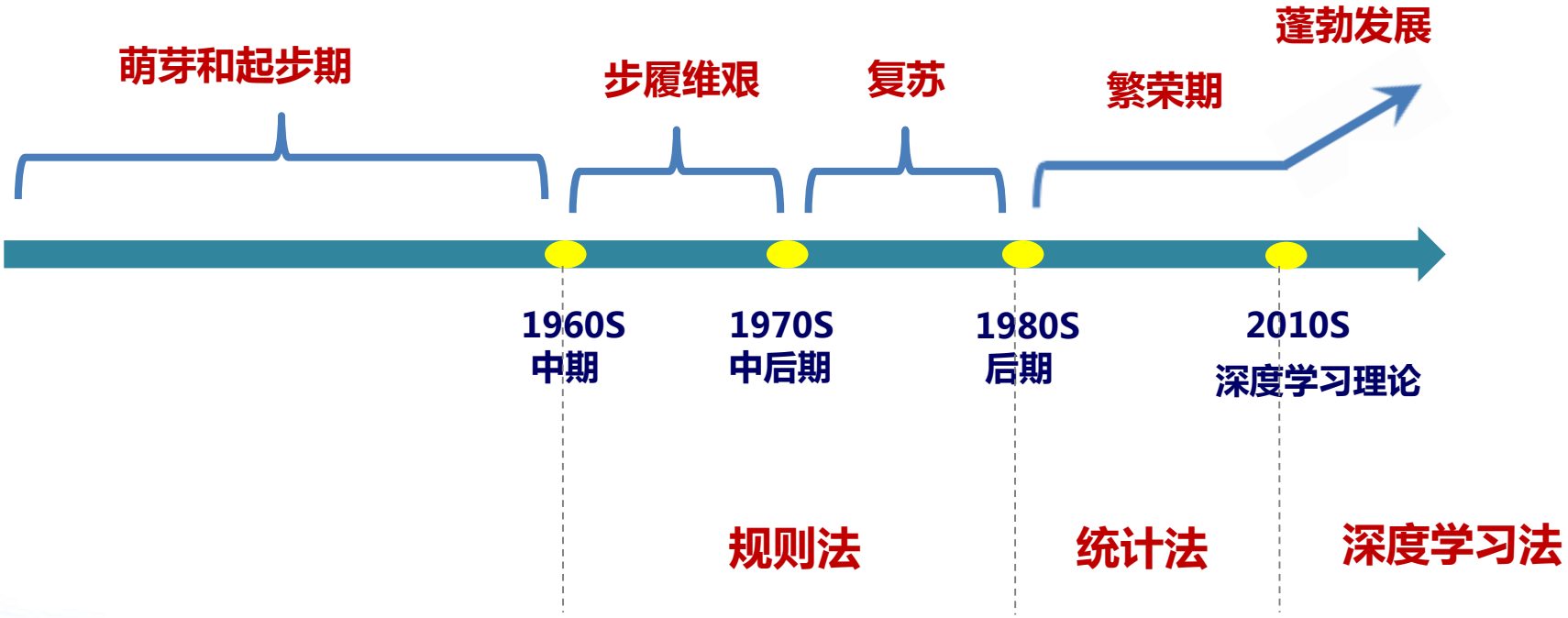
经验主义

2010s

深度学习理论（神经网络）

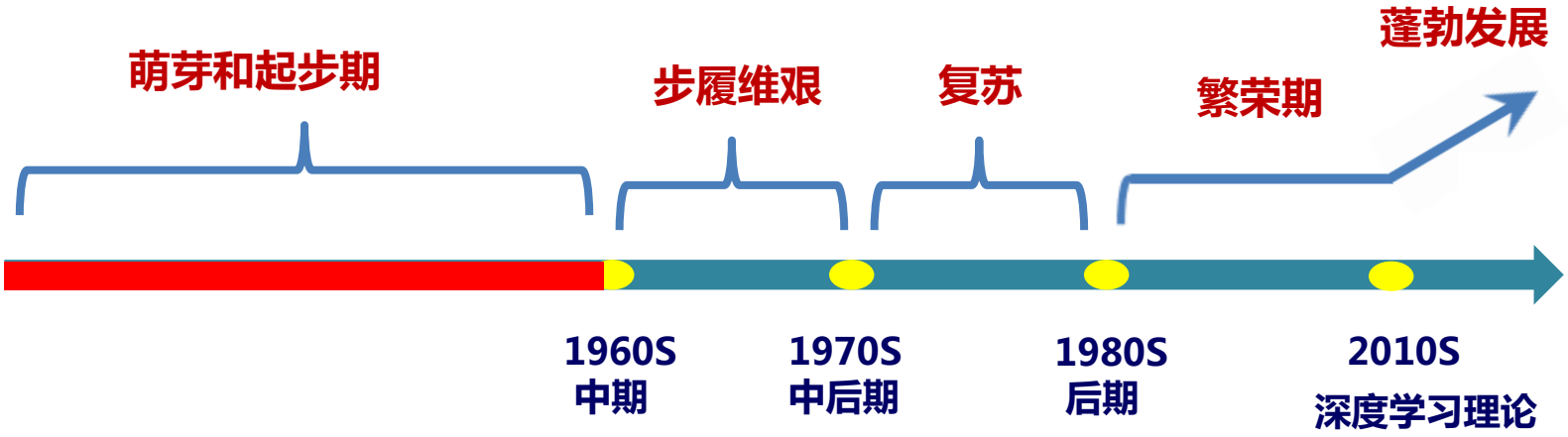
2. 自然语言处理发展历史及学派

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：



17世纪的“普遍语言”的运动（笛卡儿（Descartes），莱布尼兹（Leibniz）维尔金斯（John Wilkins）等 ），用数学方法研究语言的先驱（俄国数学家B. Buljakovski ，英国数学家A. De Morgen ，德国学者F.W. Kaeding 等）

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

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。“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。

提出“图灵测试”方法。

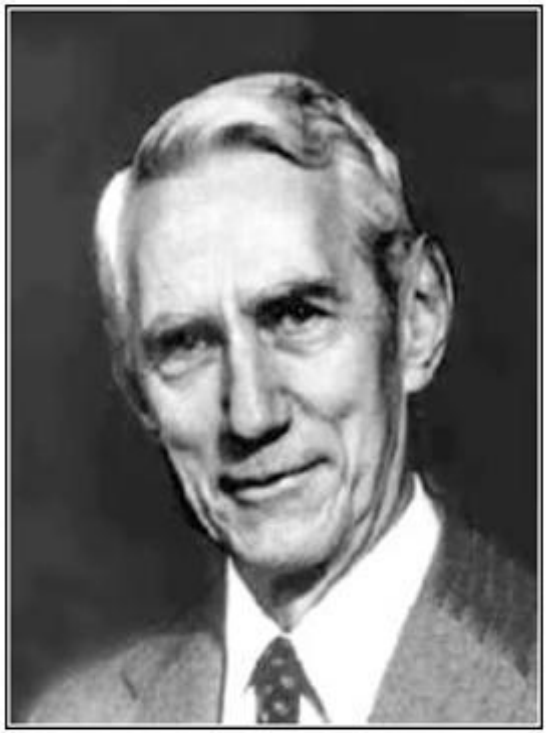
Turing

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--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

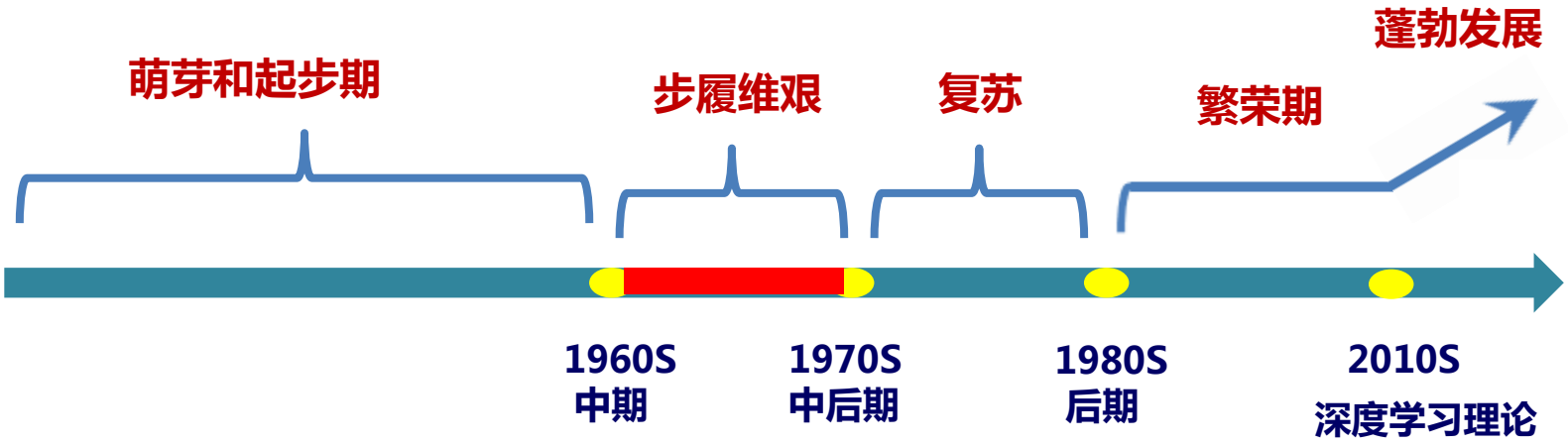
1956年，美国语言学家N. Chomsky（乔姆斯基）从Shannon的工作中吸取了有限状态马尔可夫过程的思想，把有限状态自动机作为一种工具来刻画语言的语法，并把有限状态语言定义为由有限状态语法生成的语言。产生了“形式语言理论”（formal language theory）

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

2. 自然语言处理发展历史及学派

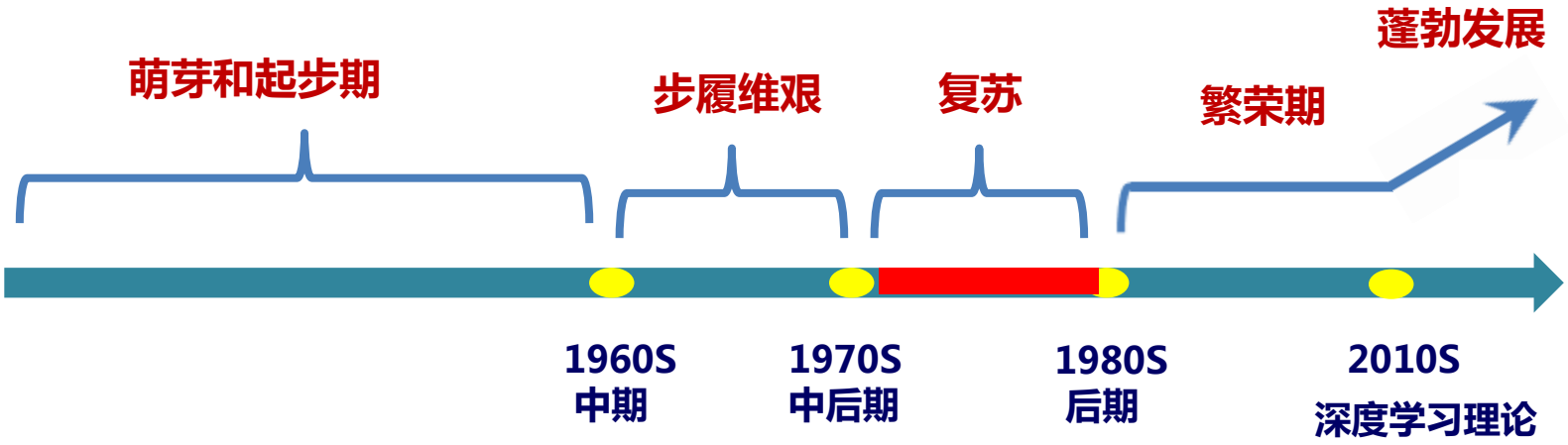
作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



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

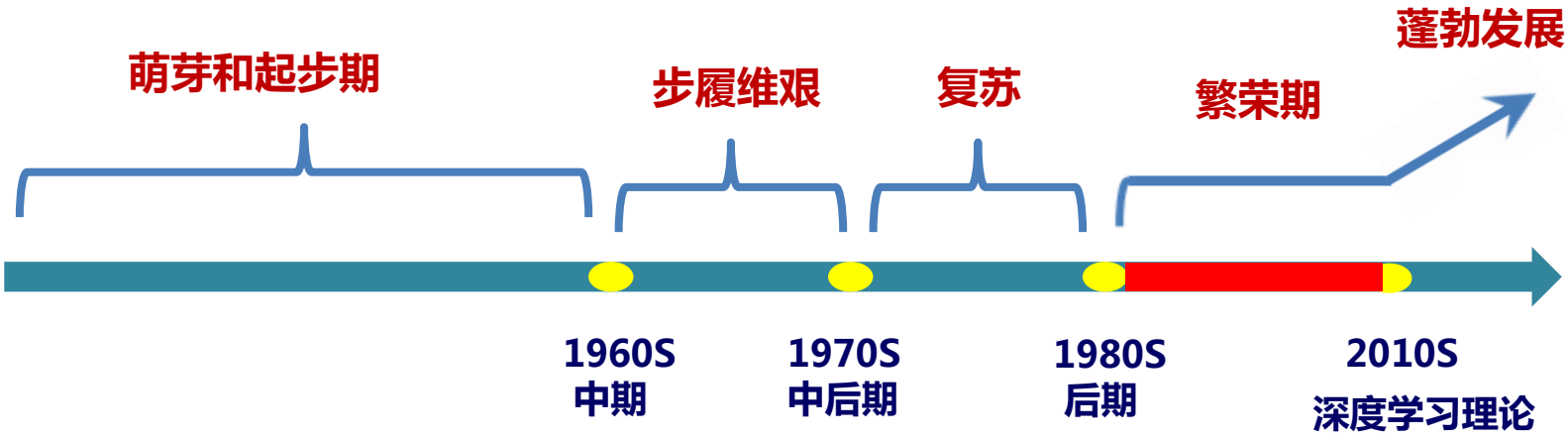
作为一门新兴的交叉学科自然语言处理经历了曲折的发展历程：



IBM的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）。

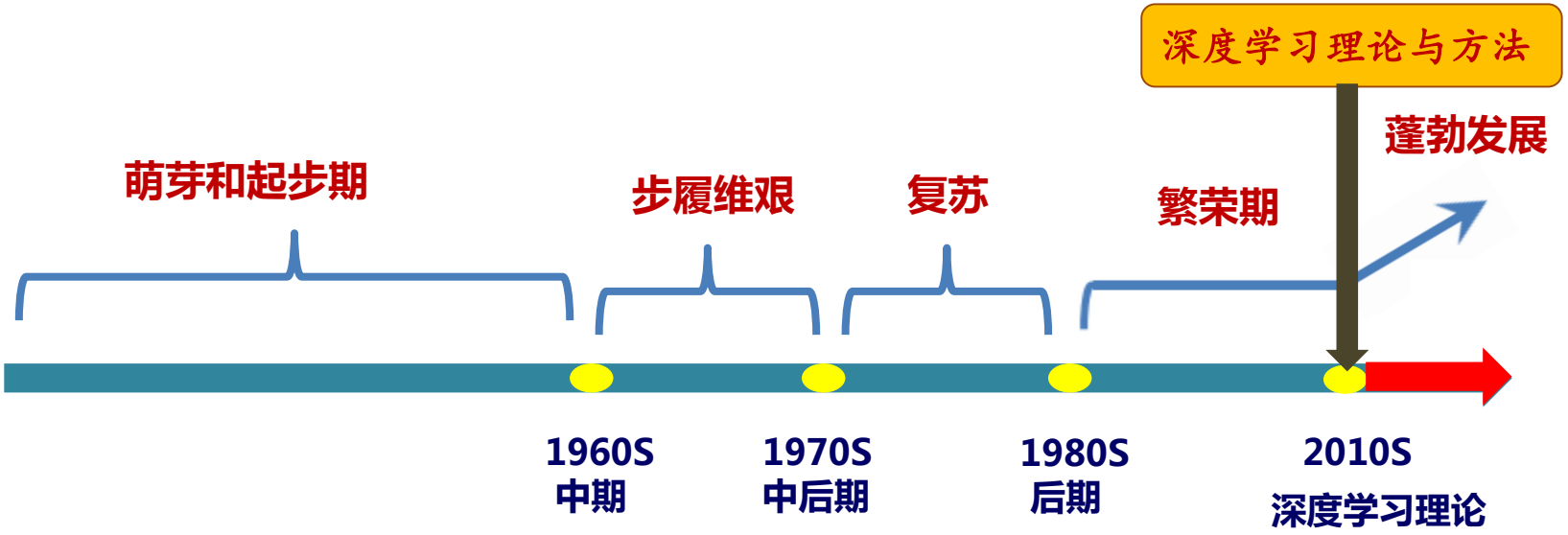
2. 自然语言处理发展历史及学派

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。

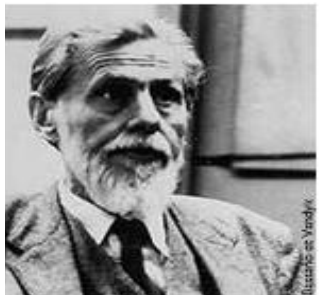
2010 S 进入了蓬勃发展（深度学习）时期



深度学习的发展历史

第一代神经网络

Walter Pitts



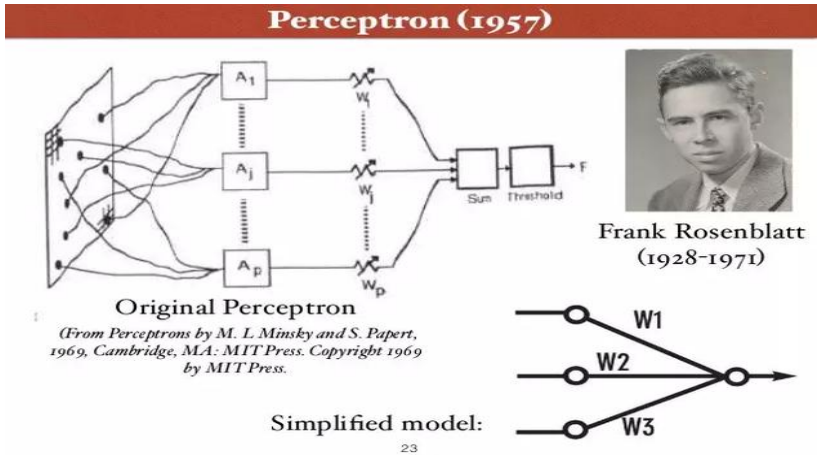
Warren McCulloch

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

深度学习的发展历史

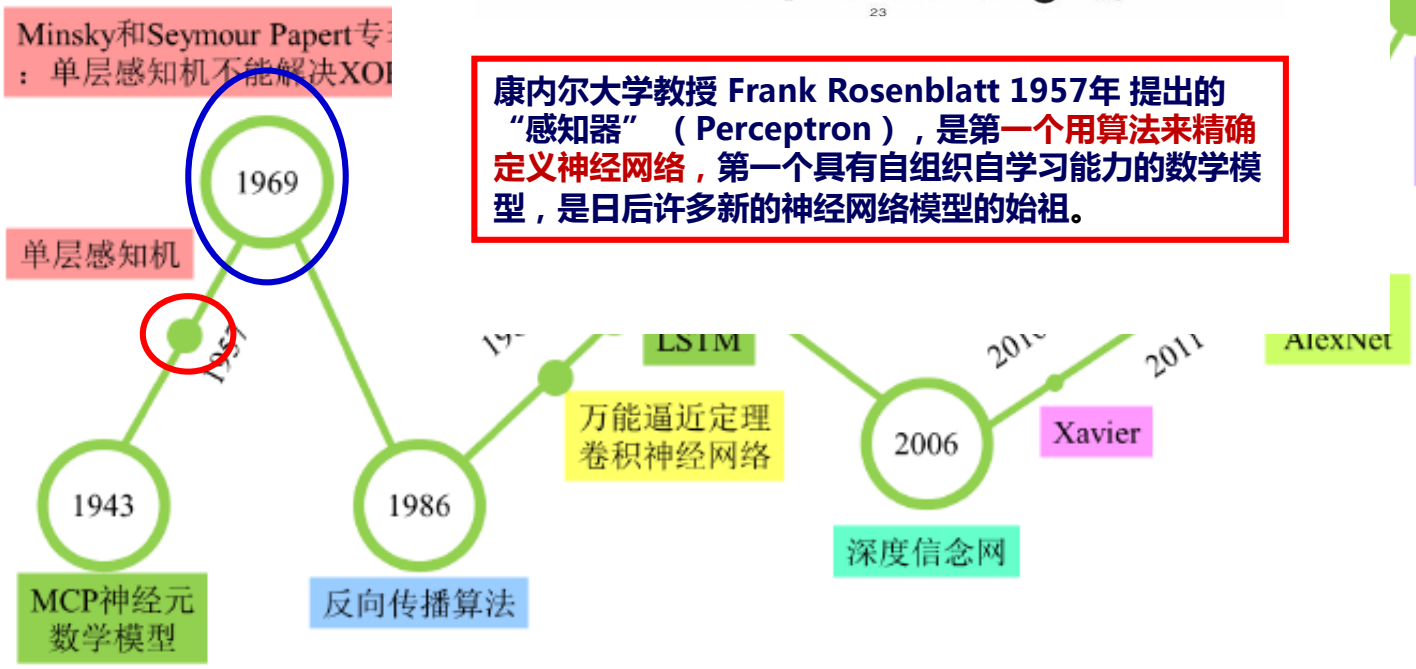
第一代

Marvin Minsky



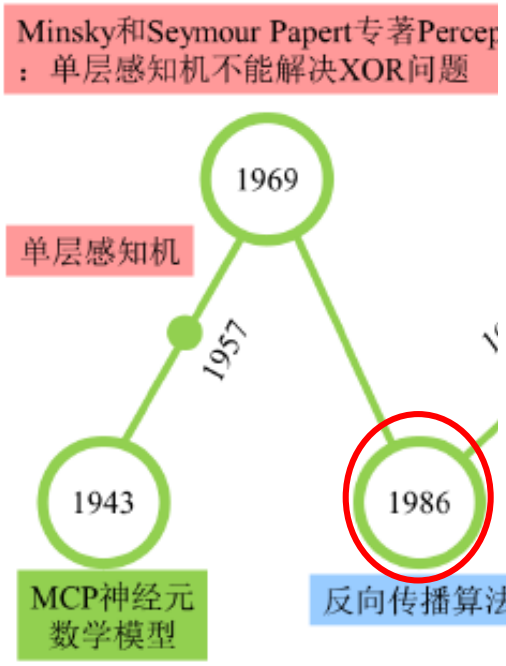
Frank Rosenblatt
(1928-1971)

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。

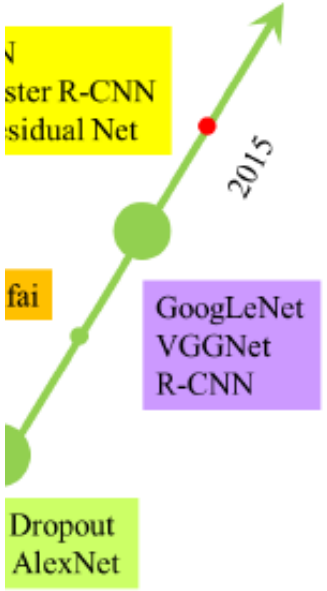


深度学习的发展历史

第二代神经网络



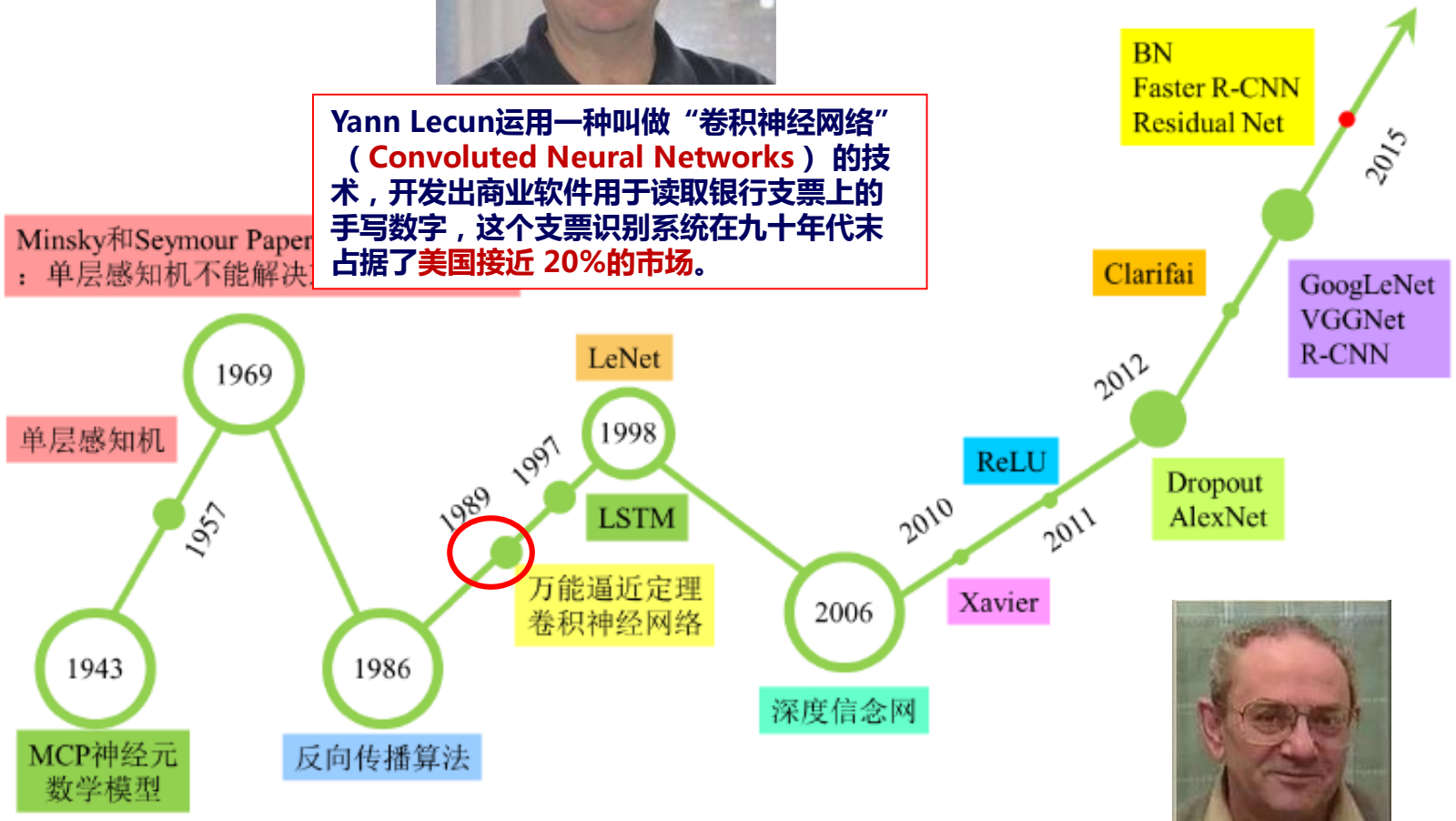
1986年 七月，Hinton 和 David Rumelhart 合作在自然杂志上发表论文，“Learning Representations by Back-propagating errors”，第一次系统简洁地阐述反向传播算法在神经网络模型上的应用。神经网络的研究开始复苏



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

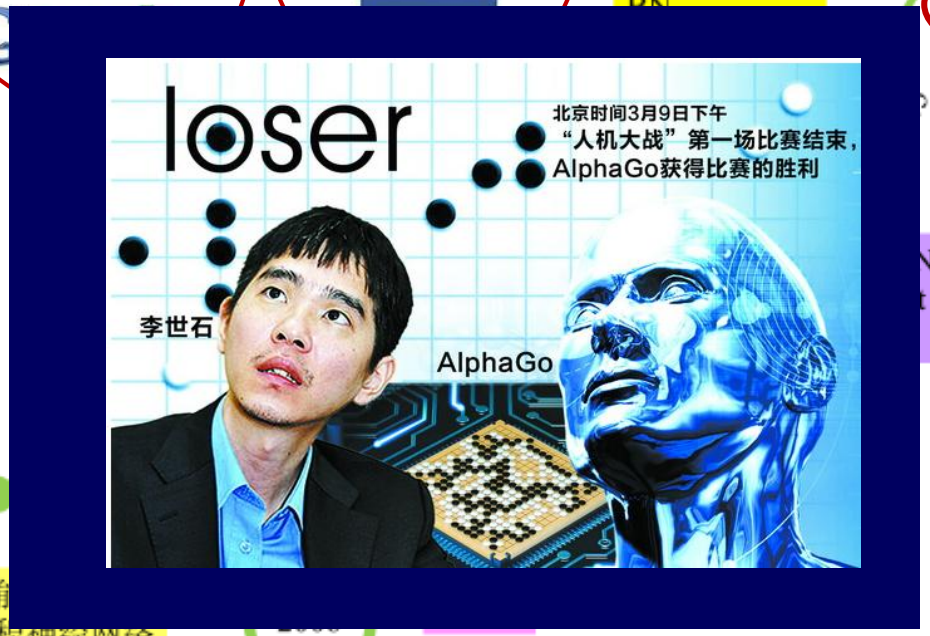
2006年，著名的学者Geoffrey Hinton在Science上发表了一篇论文，给出了训练深层网络的新思路（无监督学习、分层预训练、新的网络结构、得名“深度”学习）

优化方法的突破是第三次NN研究浪潮兴起的钥匙

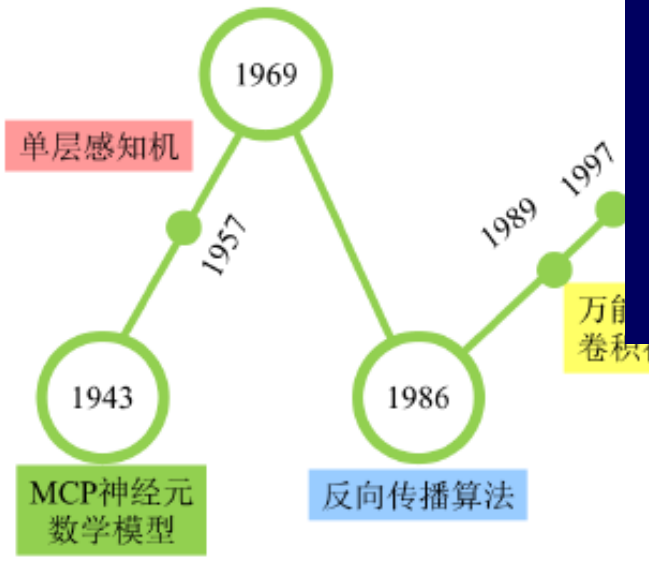
2012年底，Geoff Hinton的博士生Alex Krizhevsky、Ilya Sutskever（他们研究深度学习时间并不长）在图片分类的竞赛ImageNet上，识别结果拿了第一名。2011年冠军的准确率(top 5精度)是74.3%。2012年，Hinton和他的学生Alex等人参赛，把准确率一下提高到84.7%。靠着深度学习震惊了机器学习领域，从此大量的研究人员开始进入这个领域，一发不可收拾。截止到2015年5月份，ImageNet数据集的精度已经达到了95%以上，某种程度上跟人的分辨能力相当了。

深度学习的发展历史

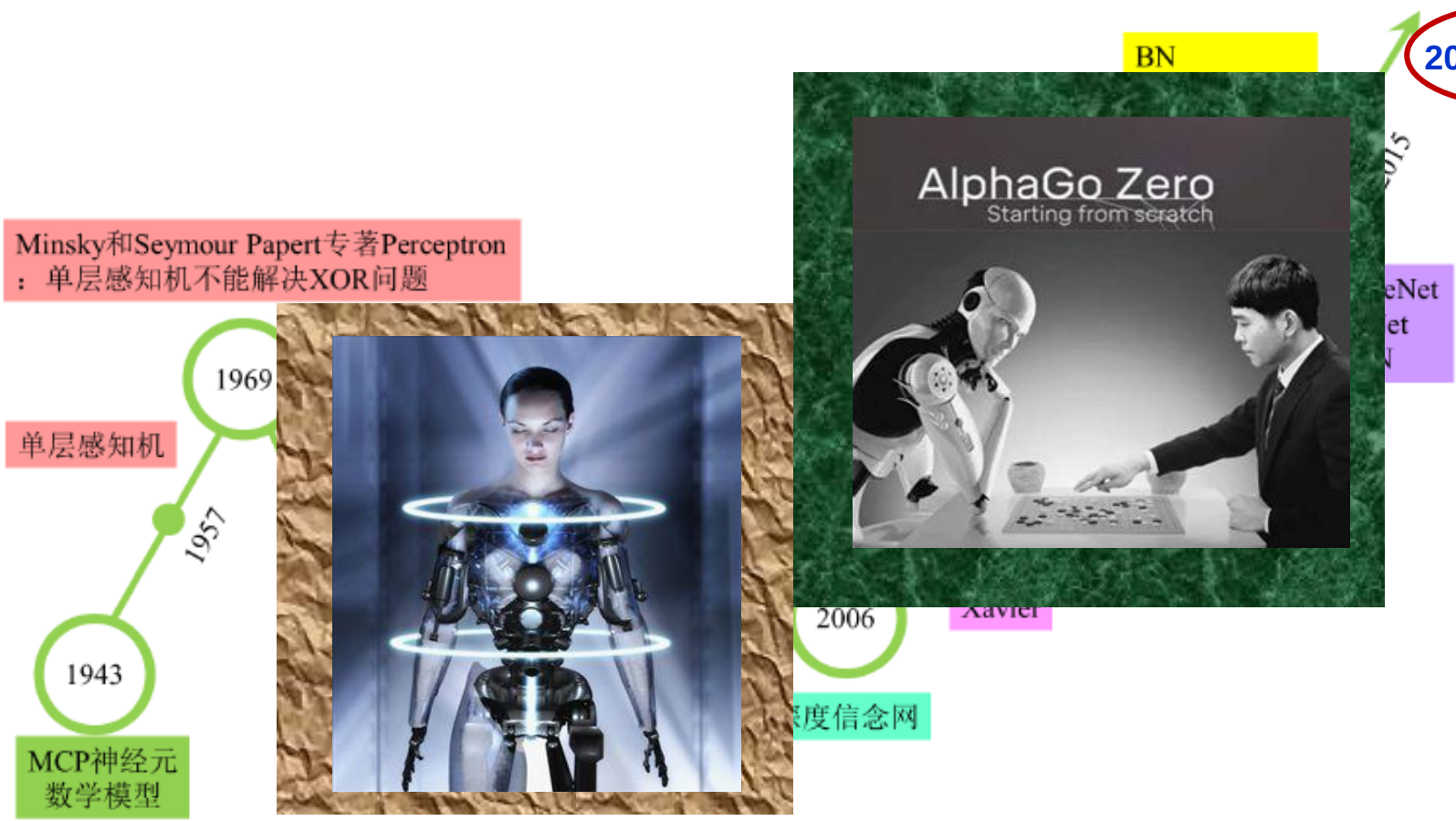
2016年3月



Minsky和Seymour Papert专著Perceptron：单层感知机不能解决XOR问题

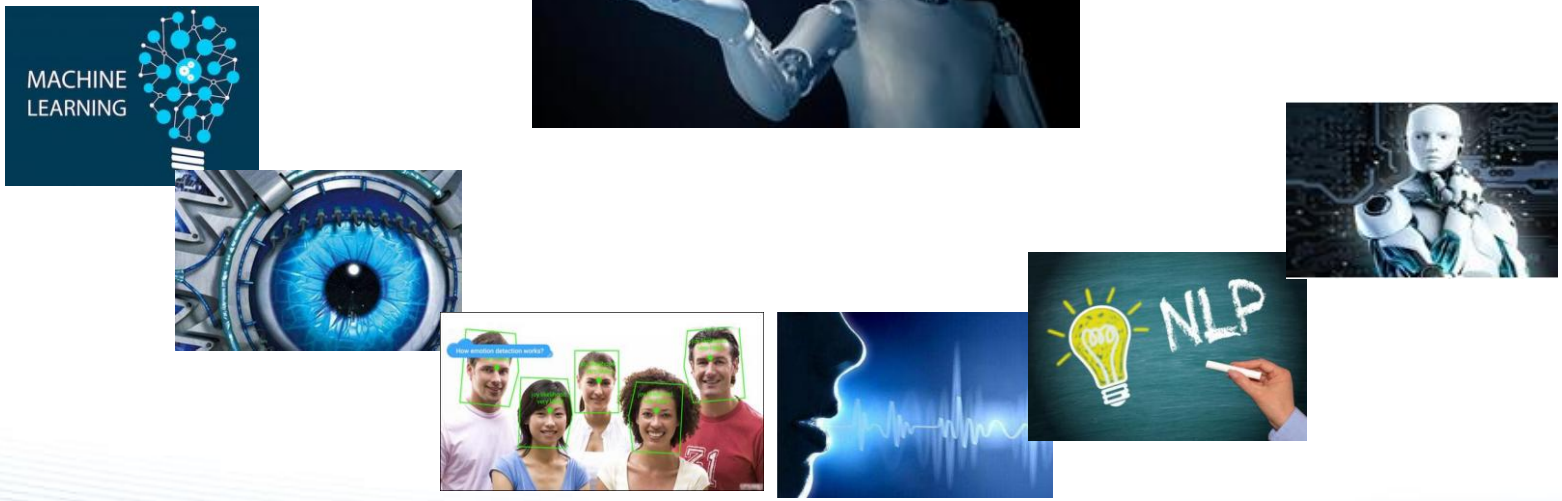


深度学习的发展历史


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
深度信念网

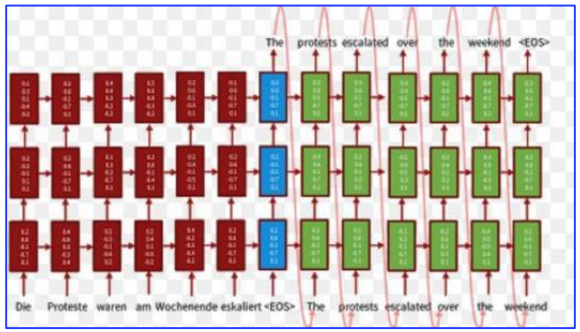
深度学习应用领域

深度学习（Deep Learning）近年来火遍了各个领域



深度学习应用领域

NLP领域



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.

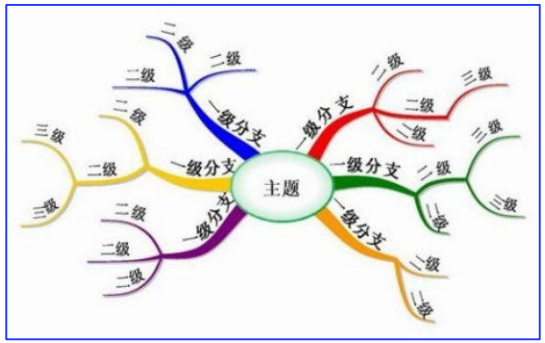
Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption in the northeast part of nigeria. he says he'll "rapidly give attention" to curbing violence in the northeast part of nigeria. he says his administration is confident it will be able to thwart criminals.

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.

深度学习应用领域

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。



2017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70%以上

内 容 提 要

1. 自然语言处理概述

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. 自然语言处理技术评测及学术会议

3 自然语言处理技术及应用架构

✧ 自然语言处理研究范围

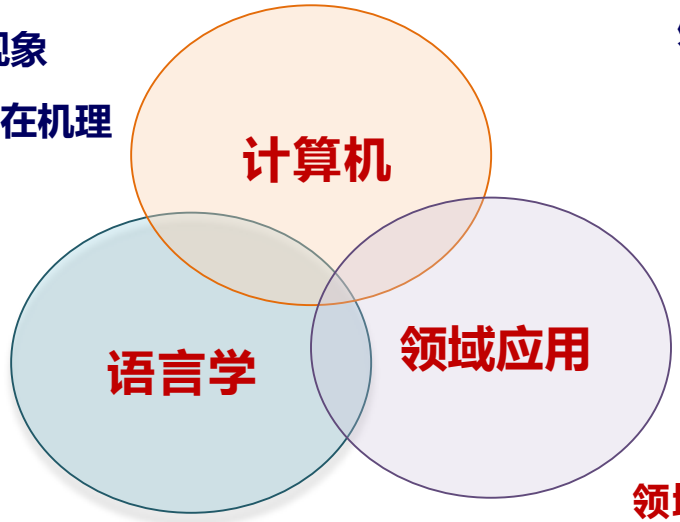
计算机 –与自然语言处理相关的计算机理论与技术

•各利用计算机对自然语言的信息进行各种处理和加工的理论、模型，技术等。

如，拼音输入、手写输入、语音识别、语言知识表示、语言分析技术 ……

语言学理论 –和解释刻画语言现象

- 人类是理解语言的内在机理
- 语言本身的内在含义
- 语言问题形式化表示



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

如，智能客服，智能教育

智能金融、智能医疗……

认知科学、语言学、逻辑学、应用数学、计算机科学等交叉学科

3 自然语言处理技术及应用架构

✧ 自然语言处理研究层面

语言是人类思维最活跃和有代表性的部分.

语言是人类交际的工具，是人类思维的载体

语言交流过程

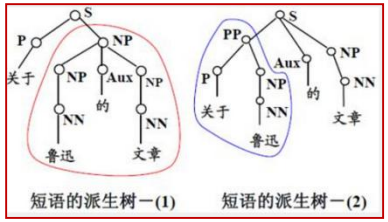
理解和消化内容的认知能力才是真正意义上的核心

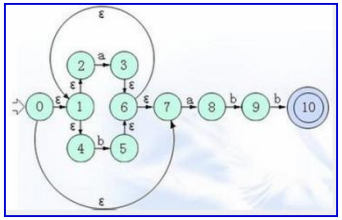
3 自然语言处理技术及应用架构

3 自然语言处理技术及应用架构

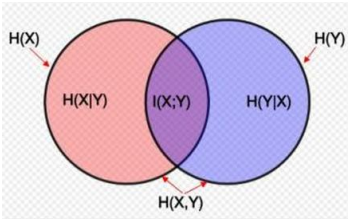
□ 基础理论

- ❖ **形式语言与自动机**
  - ➢ 形式语言与自动机



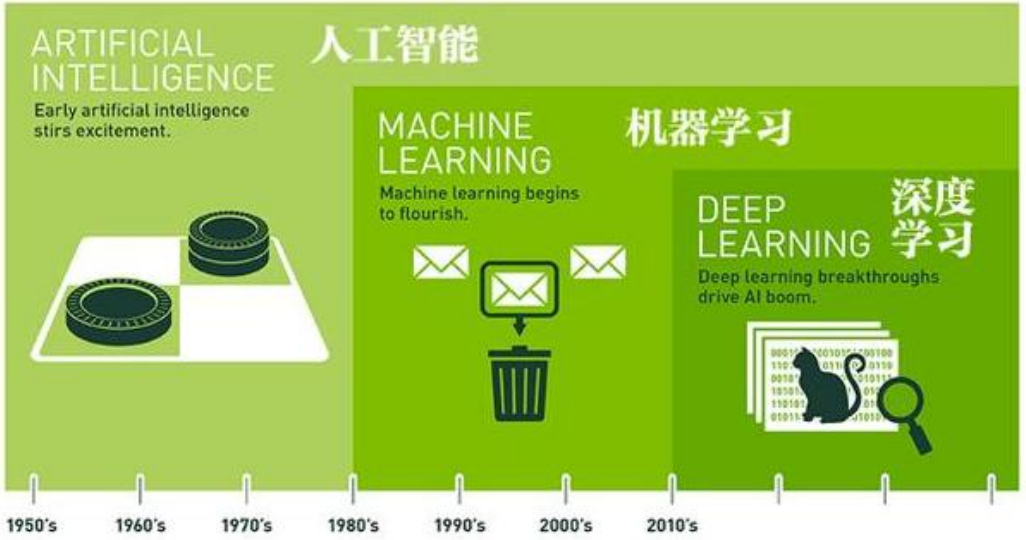


- ❖ **概率论&信息论**
  - ➢ 概率论
  - ➢ 信息论



□ 基础理论

- ❖ **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

口 数据资源

❖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，主要用于研究自然语言的规律，特别是统计语言学模型的训练以及相关系统的评价和测试。语料库是统计NLP的知识来源语料库（详细介绍见 第 2 章 ）

口 数据资源

❖语言知识库（corpus base）：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。Wordnet, Hownet 等

❖知识图谱（knowledge graph ）：把复杂的知识领域通过数据挖掘信、信息处理、知识计量和图形绘制而显示出来，揭示知识领域的动态发展规律，为学科研究提供切实的、有价值的参考。

3 自然语言处理技术及应用架构

□ NLP基础技术

- **词法分析**：词法分析目的是从句子中分出单词，找出词汇的各个词素，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

国际计算语言联合会（ACL）下设的汉语特别兴趣（SIGHAN）研究组每年举办国际汉语分词评测大赛。

□ NLP基础技术

- **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。一个实用的、完备的、准确的句法分析将是计算机真正理解自然语言的基础。
  - **短语结构分析 （主要采用宾州树库）**

    早期集中于短语结构分析方面，Collins和Charniak的工作最为著名，方法：词汇化的概率短语结构语法（Lexicalized PCFG）
  - **依存分析（图的分析算法和基于转换的分析算法）**

    CONLL会议多次将依存分析作为共享任务（Shared Task）评测

□ 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
  - **词义排歧和语义归纳、推理等（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）**

    语义角色标注属于浅层语义分析技术，其目的是为句子中的每个动词标注出其相关的名词及其语义角色

□ 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响

  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。

  “语境”的范围

  - 一段话
  - 整篇文章
  - 作者的身份和处境
  - 文化背景

□ NLP基础技术

- **篇章分析**：指超越单个句子范围的各种可能分析，包括句子（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。
  - **逻辑语义结构**

    表征并列、转折、因果等逻辑关系
  - **指代结构**

    表征名词、名词短语、代词、零形式相互之间的共指关系
  - **话题结构**

    宏观话题结构

    微观话题结构

3 自然语言处理技术及应用架构

□ NLP核心应用

按照应用目标划分，广义上包括：

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

▸应用：文献翻译、网页辅助浏览等。

▸代表系统：

- **Google：http://translate.google.cn (64 种语言)**
- **Systran：http://www.systransoft.com (15 种语言)**
- **百度： http://fanyi.baidu.com/ (汉英、汉日)**
- **有道： http://fanyi.youdao.com/ (英,日,韩,法←→汉)**

□ NLP核心应用

❖ 信息检索 (Information retrieval)

信息检索也称情报检索，就是利用计算机系统从大量文档中找到符合用户需要的相关信息。

▸代表系统：Google: http://www.google.com

百度： http://www.baidu.com.cn/

目前至少有300多亿个网页，每天数以万计地增加，只有1%的信息被有效地利用。

□ NLP核心应用

❖ 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。

实体关系抽取 (entity relation extraction)。

社会网络 (social network)

❖ 自动文摘 (Automatic summarization / Automatic abstracting)

将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。

观点挖掘 (Opinion mining) 。

▶ 应用：电子图书管理、情报获取等。

□ NLP核心应用

❖ **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。问答技术有时与语音技术和多模态输入/输出技术，以及人机交互技术等相结合，构成人机对话系统 (man-computer dialogue system)。

社区问答(Community Question Answering, CQA)

❖ **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。同时，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。

□ NLP核心应用

❖ **文档分类 (Document categorization)**

文档分类也叫文本自动分类 (Text categorization / classification) **或信息分类，其目的就是利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。**

❖ **情感分类**(Sentimental classification)

狭义的情感分析(sentiment analysis)是指利用计算机实现对文本数据的观点、情感、态度、情绪等的分析挖掘。广义的情感分析则包括对图像视频、语音、文本等多模态信息的情感计算。

□ NLP核心应用

❖ 信息推荐与过滤(formation Recommendation and Filtering)

信息推荐与过滤（简称信息推荐）是根据用户的习惯、偏好或兴趣，从大规模信息中识别满足用户兴趣的信息的过程。

3 自然语言处理技术及应用架构

□ NLP+应用

❖与领域深度结合，为行业创造价值

将自然语言处理技术深入到各个应用系统和垂直领域中。

如，银行、医药、司法、教育、金融等领域

- 应用：搜索引擎、智能客服、商业智能和语音助手、法律助手

  智能医疗系统、智能教育……

代表产品：

个人助理（闲聊型）：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）、阿里小蜜等

内 容 提 要

1. 自然语言处理概述

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. 自然语言处理技术评测及学术会议

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）
- Coling （ International Conference on Computational Linguistics）
- EMNLP （ Conference on Empirical Methods in Natural language Processing）
- 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)

详见 课本 “自然语言处理及其相关领域的国际会议”

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

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

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

练习与思考

阅读“中文信息处理发展报告，2016” 了解中文信息处理领域的基本概念、前沿技术、应用方向 和 发展趋势

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

在 此 表 示 感 谢 ！

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

谢谢各位！