自然语言处理 NLP 技术里程碑、知识结构、研究方向和机构导师(公号回复“NLP 总结”下载彩标 PDF 典藏版资料)

秦陇纪

简介：自然语言处理 NLP 技术里程碑、知识结构、研究方向和机构导师。(公号回复“NLP 总结”，文末“阅读原文”可下载 16 图 1 表 36k 字 29 页 PDF 资料)蓝色链接“数据简化 DataSimp”关注后下方菜单有文章分类页。作者：秦陇纪。来源：知网、谷歌、百科、知乎等，数据简化社区秦陇纪微信群聊公众号，引文出处附参考文献。主编译者：秦陇纪，数据简化、科学 Sciences、知识简化新媒体创立者，数据简化社区创始人 OS 架构师/C/Java/Python/Prolog 程序员，IT 教师。每天大量中英文阅读/设计开发调试/文章汇译编简化，时间精力人力有限，欢迎转发/赞赏/加入支持社区。

目录

自然语言处理 NLP 技术里程碑、知识结构、研究方向和机构导师(29847 字)

数据简化DataSimp导读：自然语言处理发展史上的十大里程碑、NLP知识结构，以及NLP国内研究方向、机构、导师。祝大家学习愉快~ 要推进人类文明，不可止步于敲门呐喊；设计空想太多，无法实现就虚度一生；工程能力至关重要，秦陇纪与君共勉之。

A 自然语言处理技术发展史十大里程碑(21585 字)

自然语言处理技术发展史十大里程碑

文|秦陇纪，参考|黄昌宁、张小凤、Sebatian Ruder，数据简化 DataSimp20181013Sat22Mon

摘要：自然语言处理(Natural Language Processing，NLP)是计算机科学、人工智能、语言学领域的学科分支、交叉学科，关注计算机和人类(自然)语言之间的相互，研究实现人与计算机之间使用自然语言进行有效通信的各种理论和方法的领域。本文从两个NLP传统研究问题出发，总结以下十大里程碑：复杂特征集、词汇主义、统计语言模型、神经语言模型、多任务学习、词嵌入、RNN/CNN用于NLP的神经网络、序列到序列模型、注意力机制网络、预训练语言模型。

关键词：自然语言处理;NLP; 中文分词; 文本分类; 信息抽取; 语义理解; 问答系统; 自然语言对话系统; 复杂特征集; 词汇主义; 语料库方法; 统计语言模型。

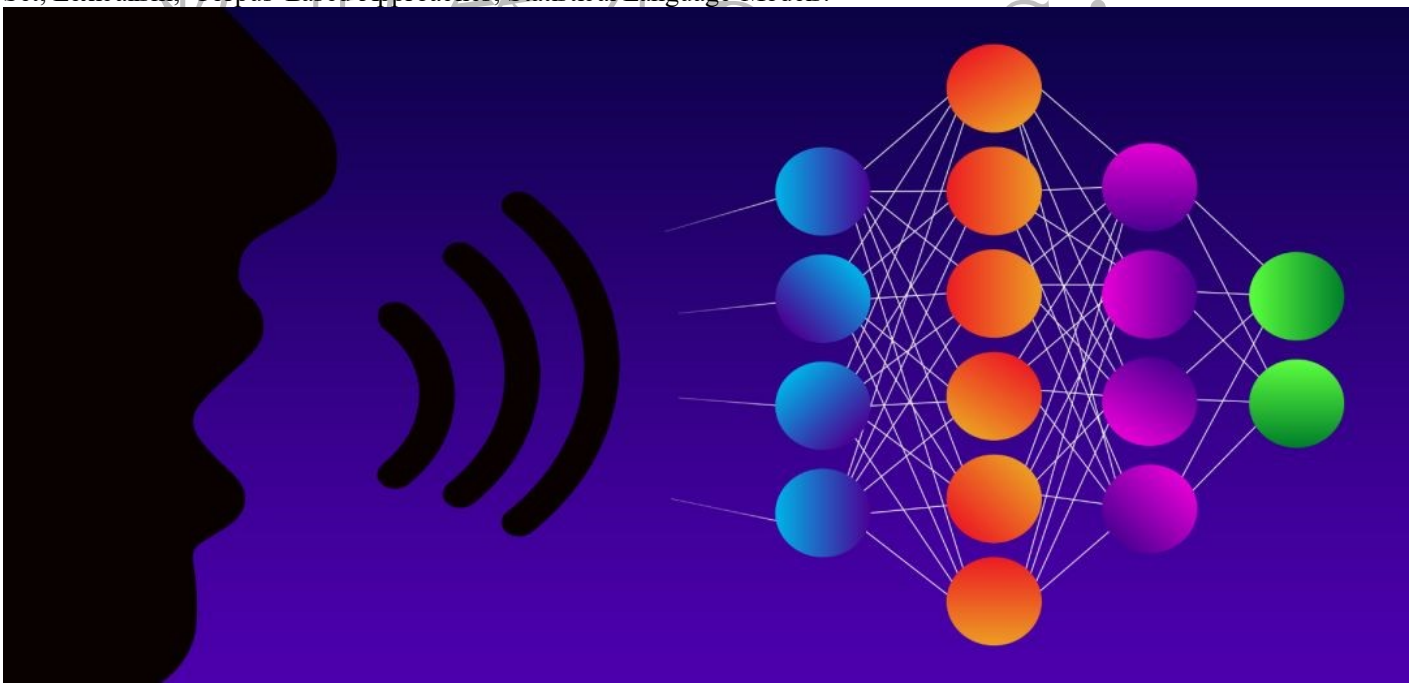
Title: **Ten milestones in the history of natural language processing technology**

Author: **Qin Longji**, data simplification Community, 20181013Sat20Sat {QinDragon2010@qq.com}

Abstract: Natural Language Processing (NLP) is a branch of interdisciplinary and interdisciplinary fields in the fields of computer science, artificial intelligence, and linguistics. It focuses on the mutual relationship between computers and human (natural) languages, and studies the use of natural language between humans and computers. The field of various theories and methods for effective communication. Based on two NLP traditional research questions, this paper summarizes the following ten major milestones: complex feature set, lexicalism, statistical language model, neural language models, .

Keywords: Natural Language Processing, NLP; Chinese Word Segmentation; Text Classification; Information Extraction; Semantic Understanding; Question Answering System; Natural Language Dialogue System; Complex Feature

Set; Lexicalism; Corpus-Based Approaches, Statistical Language Models.



自然语言是人类独有的智慧结晶。自然语言处理(Natural Language Processing，NLP)是计算机科学领域与人工智能领域中的一个重要方向，旨在研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。用自然语言与计算机进行通信，有着十分重要的实际应用意义，也有着革命性的理论意义。由于理解自然语言，需要关于外在世界的广泛知识以及运用操作这些知识的能力，所以自然语言处理，也被视为解决**人工智能完备(AI-complete)**的核心问题之一。对自然语言处理的研究也是充满魅力和挑战的。

微软亚洲研究院**黄昌宁、张小凤**在2013年发表论文，就过去50年以来自然语言处理(NLP)研究领域中的发现和发展要点进行阐述，其中包括两个事实和三大重要成果。近年来，自然语言处理的语料库调查显示如下两个事实：(1)对于句法分析来说，基于单一标记的短语结构规则是不充分的；单个标记的PSG规则不足以进行自然语言描述；(2)PSG规则在文本语料库中具有偏差分布，即PSG规则的总数似乎不能够涵盖大型语料库中发现的语言现象，这不符合语言学家的期望。短语结构规则在真实文本中的分布呈现严重扭曲。换言之，有限数目的短语结构规则不能覆盖大规模语料中的语法现象。这与原先人们的预期大相径庭。

NLP技术发展历程在很大程度上受到以上两个事实的影响，在该领域中可以称得上里程碑式的成果有如下三个：(1)复杂特征集和合一语法；(2)语言学研究中的词汇主义；(3)语料库方法和统计语言模型。业内人士普遍认为，大规模语言知识的开发和自动获取是NLP技术的瓶颈问题。因此，语料库建设和统计学习理论将成为该领域中的关键课题。

Natural language is the unique wisdom of mankind. Natural Language Processing (NLP) is an important direction in the field of computer science and artificial intelligence. It aims to study various theories and methods that can realize effective communication between human and computer in natural language. Communicating with computers in natural language has very important practical application significance and revolutionary theoretical significance. Because of the understanding of natural language, the need for extensive knowledge of the external world and the ability to manipulate it, natural language processing is also seen as one of the core issues in solving **AI-complete**. The study of natural language processing is also full of charm and challenge.

Huang Changning and Zhang Xiaofeng in Microsoft Asia Research Institute published a paper in 2013, elaborated on the major findings and developments points in the research field of Natural Language Processing (NLP) in the past 50 year, including two facts and three important achievements. In recent years the corpus investigation of Natural Language Processing as shown the following two facts: (1) For syntactic analysis, the rule structure of phrase based on single mark is not sufficient; Single labeled PSGrules are not sufficient for natural language description, and (2) PSGrules have skew distribution in text corpora, i.e. the total number of PSGrules does not seem to be able to cover the language phenomena found in a large corpus, which is out of most linguists expectation. The distribution of phrase structure rules in real text is seriously distorted. In other words, a limited number of phrase structure rules cannot cover grammatical phenomena in large-scale corpus. This is very different from the expectations of the original people.

The development of NLP technology has been under the influence of the two facts mentioned above. There have been three major breakthroughs and milestones in this field: (1) multiple features and unification-based grammars, (2) lexicalism in linguistics research, (3) Statistical Language Modeling (SLM) and corpus-based approaches. The latest investigations reveal that the bottleneck problem in the NLP technology is the problem of obtaining and developing large scale linguistic

knowledge; therefore, the corpus construction and statistical learning theory become key issues in NLPresearch and application.

一、NLP 研究传统问题

自然语言处理(NLP)是计算机科学、信息工程和人工智能的子领域，涉及计算机和人类(自然)语言之间的交互。，尤其是编程实现计算机处理和分析大量自然语言数据。自然语言处理的挑战包括语音识别，自然语言理解和自然语言生成。Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

信息输入、检索、人机对话等需求增多，使**自然语言处理(NLP)**成为21世纪初的热门学科。从50年代机器翻译和人工智能研究算起，NLP至今有长达半个世纪的历史了。近年来这一领域中里程碑式的理论和方法贡献有如下三个：(1)复杂特征集和合一语法；(2)语言学研究中的词汇主义；(3)语料库方法和统计语言模型。这三个成果将继续对语言学、计算语言学和NLP的研究产生深远影响。[21]为了理解这些成果的意义，先介绍一下两个相关事实。

句法分析的全过程：自然语言处理中识别句子的句法结构，要把**句子**中的**词**一个一个地切分出来：然后去查词典，给句子中的每个词指派一个合适的**词性(part of speech)**；之后再用句法规则把句子里包含的**句法成分**，如名词短语、动词短语、小句等，逐个地识别出来。进而判断每个短语的**句法功能**，如主语、谓语、宾语等，及其语义角色，最终得到**句子的意义表示**，如逻辑语义表达式。

1.1 事实一：语言的结构歧义问题

第一个事实(黄昌宁，张小凤，2013)是：短语结构语法(Phrase Structure Grammar，简称PSG)不能有效地描写自然语言。PSG在**Chomsky**的语言学理论[1]中占有重要地位，并且在自然语言的句法描写中担当着举足轻重的角色。但是它有一些根本性的弱点，主要表现为它使用的是像词类和短语类那样的单一标记，因此不能有效地指明和解释自然语言中的**结构歧义问题**。

让我们先来看一看汉语中“V+N”组合。假如我们把“打击，委托，调查”等词指派为**动词(V)**；把“力度，方式，盗版，甲方”等词视为**名词(N)**。而且同意“打击力度”、“委托方式”是**名词短语(NP)**,“打击盗版”、“委托甲方”是**动词短语(VP)**。那么就会产生如下两条有歧义的句法规则：

(1) NP → V N
(2) VP → V N

换句话讲，当计算机观察到文本中相邻出现的“V+N”词类序列时，仍不能确定它们组成的究竟是NP还是VP。我们把这样的歧义叫做“短语类型歧义”。例如：

- 该公司正在招聘[销售V人员N]NP。
- 地球在不断[改变V形状N]VP。

下面再来看“N+V”的组合，也同样会产生带有短语类型歧义的规则对，如：

(3) NP → N V 例：市场调查；政治影响。
(4) S → N V 例：价格攀升；局势稳定。

其中标记**S代表小句**。

不仅如此，有时当机器观察到相邻出现的“N+V”词类序列时，甚至不能判断它们是不是在同一个短语中。也就是说，“N+V”词类序列可能组成名词短语NP或小句S，也有可能根本就不在同一个短语里。后面这种歧义称为“短语边界歧义”。下面是两个相关的例句：

- 中国的[铁路N建设V]NP发展很快。
- [中国的铁路N]NP建设V得很快。

前一个例句中，“铁路建设”组成一个NP；而在后一个例句中，这两个相邻的词却分属于两个不同的短语。这足以说明，基于单一标记的PSG不能充分地描述自然语言中的句法歧义现象。下面让我们再来看一些这样的例子。

(5)NP → V N1 de N2
(6)VP → V N1 de N2

其中de代表结构助词“的”。例如，“[削苹果]VP的刀”是NP；而“削[苹果的皮]NP”则是VP。这里既有短语类型歧义，又有短语边界歧义。比如，“削V苹果N”这两个相邻的词，可能构成一个VP，也可能分处于两个相邻的短语中。

(7)NP → P N1 de N2

(8)PP → P N1 de N2

规则中P和PP分别表示介词和介词短语。例如，“[对上海]PP的印象”是NP；而“对[上海的学生]NP”则是PP。相邻词“对P 上海N”可能组成一个PP，也可能分处于两个短语中。

(9)NP → NumP N1 de N2

其中NumP 表示数量短语。规则(9)虽然表示的是一个NP，但可分别代表两种结构意义：

(9a)NumP [N1 de N2]NP 如：五个[公司的职员]NP

(9b)[NumP N1]NP de N2 如：[五个公司]NP 的职员

(10)NP → N1 N2 N3

规则(10)表示的也是一个NP，但“N1+ N2”先结合，还是“N2 +N3”先结合，会出现两种不同的结构方式和意义，即：

(10a)[N1 N2]NP N3 如：[现代汉语]NP 词典

(10b)N1 [N2 N3]NP 如：新版[汉语词典]NP

以上讨论的第一个事实说明：

！由于约束力不够，单一标记的PSG规则不能充分消解短语类型和短语边界的歧义。用数学的语言来讲，PSG规则是必要的，却不是充分的。因此机器仅仅根据规则右边的一个词类序列来判断它是不是一个短语，或者是什么短语，其实都有某种不确定性。

！采用复杂特征集和词汇主义方法来重建自然语言的语法系统，是近二十年来全球语言学界就此作出的最重要的努力。

1.2 事实二：词频统计的齐夫律

通过大规模语料的调查，人们发现一种语言的短语规则的分布也符合所谓的**齐夫率(Zipf's Law)**。**Zipf**是一个统计学家和语言学家。他提出，如果对某个**语言单位(不论是英语的字母或词)**进行统计，把这个语言单位在一个语料库里出现的**频度(frequency)**记作F，而且根据频度的降序对每个单元指派一个整数的**阶次(rank)** R。结果发现R和F的乘积近似为一个**常数**。即

F*R ≈ const (常数)

被观察的语言单元的阶次R与其频度F成反比关系。词频统计方面齐夫律显示，不管被考察的语料仅仅一本长篇小说，还是一个大规模的语料库，最常出现的100个词的出现次数会占到语料库总词**次数(tokens)**的近一半。假如语料库的规模是100万词次，那么其中频度最高的100个词的累计出现次数大概是50万词次。如果整个语料库含有5万**词型(types)**，那么其中的一半(也就是2.5万条左右)在该语料库中只出现过一次。即使把语料库的规模加大十倍，变成1000万词次，统计规律大体不变。

有趣的是，80年代英国人**Sampson**对英语语料库中的PSG规则进行统计，发现它们的分布同样是扭曲的，大体表现为齐夫率[4]。也就是说，一方面经常遇到的语法规则只有几十条左右，它们的出现频度非常非常高；另一方面，规则库中大约一半左右的规则在语料库中只出现过一次。随着语料库规模的扩大，新的规则仍不断呈现。Noam **Chomsky**曾提出过这样的假设，认为对一种自然语言来说，其语法规则的数目总是有限的，但据此生成的句子数目却是无限的。但语料库调查的结果不是这个样子。这个发现至少说明，单纯依靠语言学家的语感来编写语法规则不可能胜任大规模真实文本处理的需求，必须寻找可以从语料库中直接获取大规模语言知识的新方法。

几十年来，NLP学界曾发表过许多灿烂成果，有词法学、语法学、语义学的，有句法分析算法的，还有众多著名的自然语言应用系统。那么究竟什么是对该领域影响最大的、里程碑式的成果呢？

二、NLP 十大里程碑

2.1 里程碑一：1985复杂特征集

复杂特征集(complex feature set)又叫做**多重属性(multiple features)描写**。语言学里，这种描写方法最早出现在语音学中。美国计算语言学家**Martin Kay**于1985年在“功能合一语法”(Functional Unification Grammar，简称FUG)新语法理论中，提出**“复杂特征集”(complex feature set)**概念。后来被**Chomsky学派**采用来扩展PSG的描写能力。

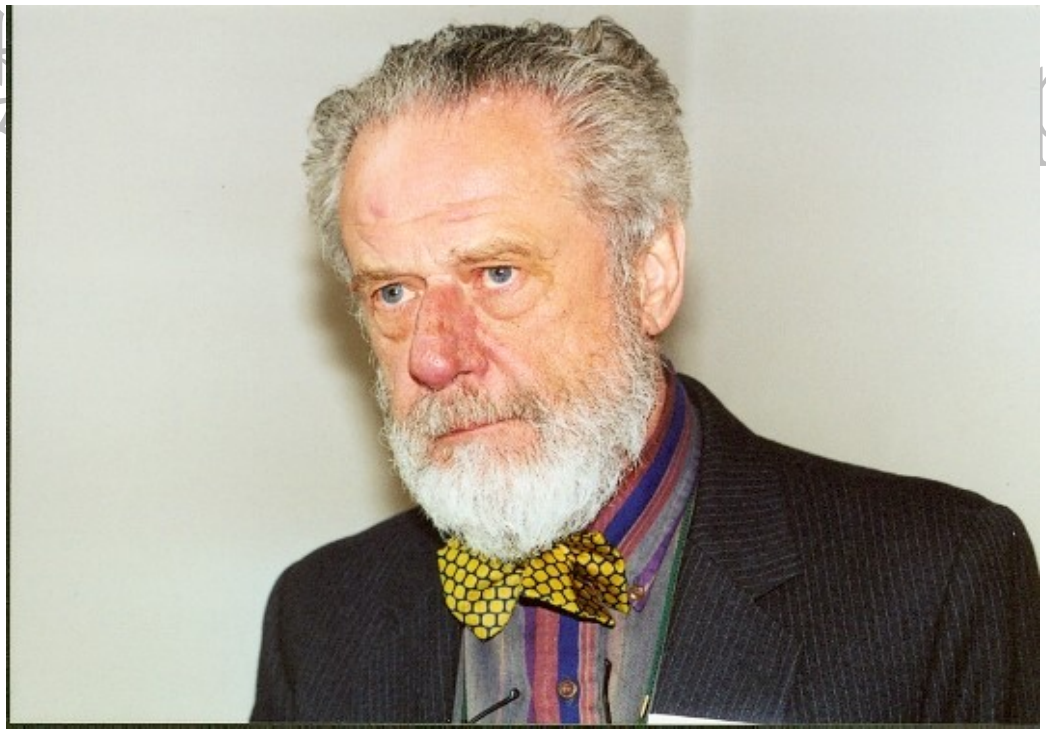


图 1 美国计算语言学家 Martin Kay

现在在语言学界、计算语言学界，语法系统在词汇层的描写中常采用复杂特征集，利用这些属性来强化句法规则的约束力。一个**复杂特征集F**包含任意多个**特征名fi**和**特征值vi**对。其形式如：

F = {…, fi=vi , …}, i=1,…,n

特征值vi既可以是一个简单的数字或符号，也可以是另外一个复杂特征集。这种递归式的定义使复杂特征集获得了强大的表现能力。举例来说，北京大学**俞士汶**开发的《现代汉语语法信息词典》[10]，对一个动词定义了约40项属性描写，对一个名词定义了约27项属性描写。

一条含有词汇和短语属性约束的句法规则具有如下的一般形式：

:<属性约束>

:<属性传递>

一般来说，**PSG规则**包括右部(条件：符号序列的匹配模式)和左部(动作：短语归并结果)。词语的“属性约束”直接来自系统的词库，而短语的“属性约束”则是在自底向上的短语归并过程中从其构成成分的**中心语(head)**那里继承过来的。在Chomsky的理论中这叫做**X-bar理论**。X-bar代表某个词类X所构成的、仍具有该词类属性的一个成分。如果X=N，就是一个具有名词特性的N-bar。当一条PSG规则的右部匹配成功，且“属性约束”部分得到满足，这条规则才能被执行。此时，规则左部所命名的的短语被生成，该短语的复杂特征集通过“属性传递”部分动态生成。

80年代末、90年代初学术界提出了一系列新的语法，如广义短语结构语法(GPSG)、中心语驱动的短语结构语法(HPSG)、词汇功能语法(LFG)等等。这些形式语法其实都是在词汇和短语的复杂特征集描写背景下产生的。合一(unification)算法则是针对复杂特征集的运算而提出来的。“合一”是实现属性匹配和赋值的一种算法，所以上述这些新语法又统称为“基于合一的语法”。

2.2 里程碑二：1966词汇主义

NLP领域第二个里程碑式贡献是**词汇主义(lexicalism)**。1966年，**韩礼德(Halliday)**提出词汇不是用来填充语法确定的一套“空位”(slots)，而是一个独立的语言学层面；词汇研究可以作为对语法理论的补充，却不是语法理论的一部分，他主张把词汇从语法研究中独立地分离出来。语言学家**Hudson**宣称，词汇主义是当今语言学理论头号发展倾向[5]。出现原因也同上节两事实有关。词汇主义方法不仅提出一种颗粒度更细的语言知识表示形式，而且体现一语言知识递增式开发和积累的新思路。

首先解释一个背景矛盾。一方面，语言学界一向认为，不划分词类就无法讲语法，如前面介绍的短语结构语法，语法“不可能”根据个别单独的词来写规则。但是另一方面，人们近来又注意到，任何归类其实都会丢失个体的某些重要信息。所以从前文提到的第一个事实出发，要想强化语法约束能力，词汇的描写应当深入到比词类更细微的词语本身上来。换句话讲，语言学呼唤在词汇层采用颗粒度更小的描写单元。从本质上来说，词汇主义倾向反映了语言描写的主体已经从句法层转移到了词汇层；这也就是所谓的“小语法，大词库”的思想。下面让我们来看与词汇主义有关的一些工作。

2.2.1 词汇语法学(Lexicon-grammar)

法国巴黎大学**Gross**教授60年代创立研究中心LADL(http://www.ladl.jussieu.fr/)，提出了词汇语法的概念。

- 把12,000个主要动词分成50个子类。
- 每个动词都有一个特定的论元集。
- 每一类动词都有一个特定的矩阵，其中每个动词都用400个不同句式来逐一描写("+"代表可进入该句式；"-"表示不能)。
- 已开发英、法、德、西等欧洲语言的大规模描写。
- INTEX是一个适用于大规模语料分析的工具，已先后被世界五十多个研究中心采用。

2.2.2 框架语义学(Frame Semantics)

Fillmore是**格语法(Case Grammar)**创始人，前几年主持美国自然科学基金的一个名为**框架语义学**的项目(http://www.icsi.berkeley.edu/~framenet)。该项目从**WordNet**上选取了2000个动词，从中得到75个语义框架。例如，动词"categorize"的框架被定义为:

一个人(Cognizer)把某个对象(Item)视为某个类(Category)。

同原先的格框架相比，原来一般化的动作主体被具体化为认知者Cognizer，动作客体被具体化为事物Item，并根据特定体动词的性质增加了一个作为分类结果的语义角色Category。

项目组还从英国国家语料库中挑出50,000个相关句子，通过人工给每个句子标注了相应的语义角色。例句：

Kim categorized the book as fiction.
(Cog) (Itm) (Cat)

2.2.3 WordNet

WordNet是一个描写英语词汇层语义关系的词库(http://www.cogsci.princeton.edu:80/~wn/)，1990年由普林斯顿大学**Miller**开发。至今有很多版本，全部公布在因特网上，供研究人员自由下载。欧洲有一个Euro-WordNet，以类似的格式来表现各种欧洲语言的词汇层语义关系。WordNet刻意描写的是词语之间的各种语义关系，如同义关系(synonymy)、反义关系(antonymy)、上下位关系(hyponymy)，部分-整体关系(part-of)等等。这种词汇语义学又叫做**关系语义学**。这一学派同传统的语义场理论和和语义属性描写理论相比，其最大的优势在于第一次在一种语言的整个词汇表上实现了词汇层的语义描写。这是其他学派从来没有做到的。其他理论迄今仅仅停留在教科书或某些学术论文中，从来就没有得到工程规模的应用。下面是WordNet的概况：

- 95,600条实词词型(动词、名词、形容词)
- 被划分成70,100个同义词集(synsets)

2.2.4 知网网(How-Net)

知网是**董振东**和**董强**[9]设计的一个汉语语义知识网(http://www.keenage.com)，访问只有主页。

- 自下而上地依据概念对汉语实词进行了穷尽的分类。
- 15,000个动词被划分成810类。
- 定义了300个名词类，100个形容词类。
- 全部概念用400个语义元语来定义。

知网特点是既有WordNet所描写的同一类词间语义关系(如：同义、反义、上下位、部分-整体等)，又描写不同类词之间的论旨关系和语义角色。

3.2.5 MindNet

MindNet是微软研究院NLP组设计的词汇语义网(http://research.microsoft.com/nlp/)，用**三元组(triple)**作为全部知识的表示基元。一个三元组由两个节点和一条连接边组成。每个节点代表一个概念，连接两个概念节点的边表示概念之间的语义依存关系。全部三元组通过句法分析器自动获取。

具体通过对两部英语词典(Longman Dictionary of Contemporary English，American Heritage Dictionary)和一部百科全书(Encarta)中的全部句子进行分析，获得每个句子的逻辑语义表示(logical form，简称LF)。而LF本来就是由三元组构成的，如(W1，V-Obj, W2)表示：W1是一个动词，W2是其宾语中的中心词，因此W2从属于W1，它们之间的关系是V-Obj。比如(play，V-Obj, basketball)便是一个具体的三元组。又如(W1, H-Mod, W2)，W1代表一个偏正短语中的**中心词(head word)**，W2是其**修饰语(modifier)**，因此W2从属于W1，它们之间的关系是H-Mod。

这种资源是完全自动做出来的，所得三元组不可能没有错误。但是那些出现频度很高的三元组一

般来说正确。MindNet已经应用到像语法检查、句法结构排歧、词义排歧、机器翻译等许多场合。

2.3 里程碑三：1976统计语言模型

第三大贡献是语料库方法，或叫统计语言模型。首先成功利用数学方法解决自然语言处理问题的是语音和语言处理大师**弗雷德·贾里尼克(Fred Jelinek)**。1968年始在IBM研究中心兼职1974年全职加入，他领导一批杰出科学家利用大型计算机处理人类语言问题。学术休假(Sabbatical Leave)时(约1972-1976年间)提出统计语言模型。1990s**李开复**用统计语言模型把997个词的语音识别问题简化成了20词识别问题，实现了有史以来第一次大词汇量非特定人连续语言的识别。常用统计语言模型，包括N元文法模型(N-gram Model)、隐马尔科夫模型(Hidden Markov Model，简称HMM)、最大熵模型(Maximum Entropy Model)等。

图 2 现代语音识别和自然语言处理研究的先驱、美国工程院院士 Frederick Jelinek

如果用**变量W**代表一个文本中顺序排列的n个词，即W = w1w2…wn，则统计语言模型的任务是给出任意一个**词序列W**在文本中出现的**概率P(W)**。利用概率的乘积公式，P(W)可展开为：

$$P(W) = P(w1)P(w2/w1)P(w3/\ w1\ w2)…P(wn/w1\ \ w2…wn\text{-}1) \quad (1)$$

式中P(w1)表示第一个词w1的出现概率，P(w2/w1)表示在w1出现的情况下第二个词w2出现的条件概率，依此类推。不难看出，为了预测词wn的出现概率，必须已知它前面所有词的出现概率。从计算上来看，这太复杂了。如果近似认为任意一个词wi的出现概率只同它紧邻的前一个词有关，那么计算就得以大大简化。这就是所谓的二元模型(bigram)，由(1)式得：

$$P(W) \approx P(w1)\ \prod i=2,…,nP(wi/\ \ wi\text{-}1\ ) \quad (2)$$

式中$\prod i=2,…,nP(wi/\ \ wi\text{-}1\ )$表示多个概率的连乘。

需要着重指出的是：这些概率参数都可以通过大规模语料库来估值。比如二元概率

$$P(wi/\ \ wi\text{-}1) \approx count(wi\text{-}1\ \ wi)\ /\ count(wi\text{-}1) \quad (3)$$

式中count(…)表示一个特定词序列在整个语料库中出现的**累计次数**。若语料库的总词次数为N，则**任意词wi**在该语料库中的出现概率可估计如下：

$$P(wi) \approx count(wi)\ /\ N \quad (4)$$

同理，如果近似认为任意词wi的出现只同它紧邻前两个词有关，就得到一个三元模型(trigram)：

$$P(W) \approx P(w1)P(w2/w1)\ \prod i=3,…,nP(wi/wi\text{-}2\ \ w\text{-}1\ ) \quad (5)$$

统计语言模型的方法有点像天气预报。用来估计概率参数的大规模语料库好比是一个地区历年积累起来的气象记录，而用三元模型来做天气预报，就像是根据前两天的天气情况来预测当天的天气。天气预报当然不可能百分之百正确。这也算是概率统计方法的一个特点。

2.3.1 语音识别

语音识别作为计算机汉字键盘输入的一种图代方式，越来越受到信息界人士的青睐。所谓听写机

就是这样的商品。据报道中国的移动电话用户已超过一亿，随着移动电话和个人数字助理(PDA)的普及，尤其是当这些随身携带的器件都可以无线上网的时候，广大用户更迫切期望通过语音识别或手写板而不是小键盘来输入简短的文字信息。

其实，语音识别任务可视为计算以下条件概率的极大值问题：

W*= argmaxW P(W/speech signal)

= argmaxW P(speech signal/W) P(W) / P(speech signal)

= argmaxW P(speech signal/W) P(W) (6)

式中数学符号argmaxW表示对不同的**候选词序列W**计算**条件概率P(W/speech signal)**的值，从而使W*成为其中条件概率值最大的那个词序列，这也就是计算机选定的识别结果。换句话讲，通过式(6)的计算，计算机找到了最适合当前输入语音信号speech signal的词串W*。

式(6)第二行是利用贝叶斯定律转写的结果，因为**条件概率P(speech signal/W)**比较容易估值。公式的**分母P(speech signa)**对给定的语音信号是一个常数，不影响极大值的计算，故可以从公式中删除。在第三行所示的结果中，P(W)就是前面所讲得统计语言模型，一般采用式(5)所示的三元模型；P(speech signal/W)叫做**声学模型**。

讲到这儿，细心的读者可能已经明白，汉语拼音输入法中的**拼音－汉字转换任务**其实也是用同样方法实现的，而且两者所用的汉语语言模型(即二元或三元模型)是同一个模型。

据笔者所知，目前市场上的听写机产品和微软拼音输入法(3.0版)都是用词的三元模型实现的，几乎完全不用句法-语义分析手段。为什么会出现这样的局面呢？这是优胜劣汰的客观规律所决定的。可比的评测结果表明，用三元模型实现的拼音-汉字转换系统，其出错率比其它产品减少约50%。

2.3.2 词性标注

一个词库中大约14%的词型具有不只一个词性。而在一个语料库中，占总词次数约30%的词具有不止一个词性。所以对一个文本中的每一个词进行词性标注，就是通过上下文的约束，实现词性歧义的消解。历史上曾经先后出现过两个**自动词性标注系统**。一个采用上下文相关的规则，叫做**TAGGIT(1971)**，另一个应用词类的二元模型，叫做**CLAWS(1987)**[2]。两个系统都分别对100万词次的英语非受限文本实施了词性标注。结果显示，采用统计语言模型的CLAWS系统的标注正确率大大高于基于规则方法的TAGGIT系统。请看下表的对比：

系统名	TAGGIT(1971)	CLAWS(1987)
标记数	86	133
方法	3000 条 CSG	
规则	隐马尔科夫模型	
标注精度	77%	96%
测试语料	布朗 LOB	

令C和W分别代表词类标记序列和词序列，则词性标注问题可视为计算以下条件概率的极大值:

C*= argmaxC P(C/W)

= argmaxC P(W/C)P(C) / P(W)

≈ argmaxC ∏i=1,…,nP(wi/ci)P(ci /ci-1) (7)

式中P(C/W)是已知输入词序列W的情况下，出现**词类标记序列C**的条件概率。数学符号argmaxC表示通过考察不同的候选词类标记序列C，来寻找使条件概率取最大值的那个词类标记序列C*。后者应当就是对W的词性标注结果。

公式第二行是利用贝叶斯定律转写的结果，由于分母P(W)对给定的W是一个常数，不影响极大值的计算，可以从公式中删除。接着对公式进行近似。首先，引入独立性假设，认为任意一个词wi的出现概率近似只同当前词的词类标记ci有关，而与周围(上下文)的词类标记无关。于是词汇概率可计算如下：

P(W/C) ≈ ∏i=1,…,n P(wi/ci) (8)

其次，采用二元假设，即近似认为任意一个**词类标记ci**的出现概率只同它紧邻的前一个词类标记ci-1有关。有

P(C) ≈ P(c1) ∏i=2,…,n P(ci /ci-1) (9)

P(ci /ci-1)是词类标记的转移概率，也叫做基于词类的二元模型。

上述这两个概率参数都可以通过带词性标记的语料库来分别估计：

P(wi/ci) ≈ count(wi,ci) / count(ci) (10)

P(ci /ci-1) ≈ count(ci-1ci) / count(ci-1) (11)

据文献报道，采用统计语言模型方法汉语和英语的次性标注正确率都可以达到96%左右[6]。

2.3.3 介词短语PP的依附歧义

英语中介词短语究竟依附于前面的名词还是前面的动词，是句法分析中常见的结构歧义问题。下例用语料库方法来解决这个问题，以及这种方法究竟能达到多高的正确率。

例句：Pierre Vinken, 61 years old, joined the board as a nonexecutive director.

令A=1表示名词依附，A=0为动词依附，则上述例句的PP依附问题可表为：

(A=0, V=joined, N1=board, P=as, N2=director)

令V, N1, N2分别代表句中动词短语、宾语短语、介宾短语的中心词，并在一个带有句法标注的语料库(又称树库)中统计如下四元组的概率Pr：

Pr = (A=1 / V=v, N1=n1, P=p, N2=n2) (10)

对输入句子进行PP 依附判断的算法如下：

若Pr = (1 / v, n1, p, n2) ≥ 0.5,

则判定PP依附于n1,

否则判定PP依附于v。

Collins和**Brooks**[8]实验使用的语料库是**宾夕法尼亚大学**标注的华尔街日报(WSJ)树库，包括：训练集20,801个四元组，测试集3,097个四元组。他们对PP依附自动判定精度的上下限作了如下分析：

一律视为名词依附(即A≡1) 59.0%

只考虑介词p的最常见附加72.2%

三位专家只根据四个中心词判断88.2%

三位专家根据全句判断93.2%

很明显，自动判断精确率的下限是72.2%，因为机器不会比只考虑句中介词p的最常见依附做得更差了；上限是88.2%，因为机器不可能比三位专家根据四个中心词作出的判断更高明。

论文报告，在被测试的3,097个四元组中，系统正确判断的四元组为2,606个，因此平均精确率为84.1%。这与上面提到的上限值88.2%相比，应该说是相当不错的结果。

传统三大技术里程碑小结

语言学家在不论是复杂特征集和合一语法，还是词汇主义方法，都是原先所谓的理性主义框架下做出的重大贡献。词汇主义方法提出了一种颗粒度更细的语言知识表示形式，而且体现了一种语言知识递增式开发和积累的新思路，值得特别推崇。尤其值得重视的是，在众多词汇资源的开发过程中，语料库和统计学习方法发挥了很大的作用。这是经验主义方法和理性主义方法相互融合的可喜开端，也是国内知名语言学者**冯志伟**等人认可的研究范式。

语料库方法和统计语言模型，国内同行中实际上存在不同评价。有种观点认为NLP必须建立在语言理解基础上，他们不大相信统计语言模型在语音识别、词性标注、信息检索等应用领域中所取得的进展。这些争论不能澄清，是因为同行间缺少统一评测。有评测才会有鉴别。

评判某方法优劣应公开、公平、相互可比的评测标准，而非研究员设计“自评”。**黄昌宁、张小凤**2013年论文表示，语料库方法和统计语言模型是当前自然语言处理技术的主流，其实用价值已在很多应用系统中得到充分证实。统计语言模型研究在结构化对象的统计建模方面，仍有广阔发展空间。自然语言处理领域业界知名博主**Sebatian Ruder**在2018年文章[22]从**神经网络**技术角度，总结**NLP领域近15年重大进展、8大里程碑事件**，提及很多神经网络模型。这些模型建立在同一时期非神经网络技术之上，如上述三大里程碑。下面接着看后续NLP技术的发展。

2.4 里程碑四：2001神经语言模型(Neural language models)

语言模型解决的是在给定已出现词语的文本中，预测下一个单词的任务。这是最简单的语言处理任务，有许多具体实际应用，如智能键盘、电子邮件回复建议等。语言模型历史由来已久，经典方法基于 **n-grams 模型**(利用前面 n 个词语预测下一个单词)，并利用平滑操作处理不可见的 n-grams。

第一个神经语言模型，**前馈神经网络**(feed-forward neural network)，是 Bengio 等人于 2001 年提出的。模型以某词语之前出现的 n 个词语作为输入向量，也就是现在大家说的词嵌入(word embeddings)向量。这些词嵌入在级联后进入一个隐藏层，该层的输出然后通过一个 softmax 层。如图 3 所示。

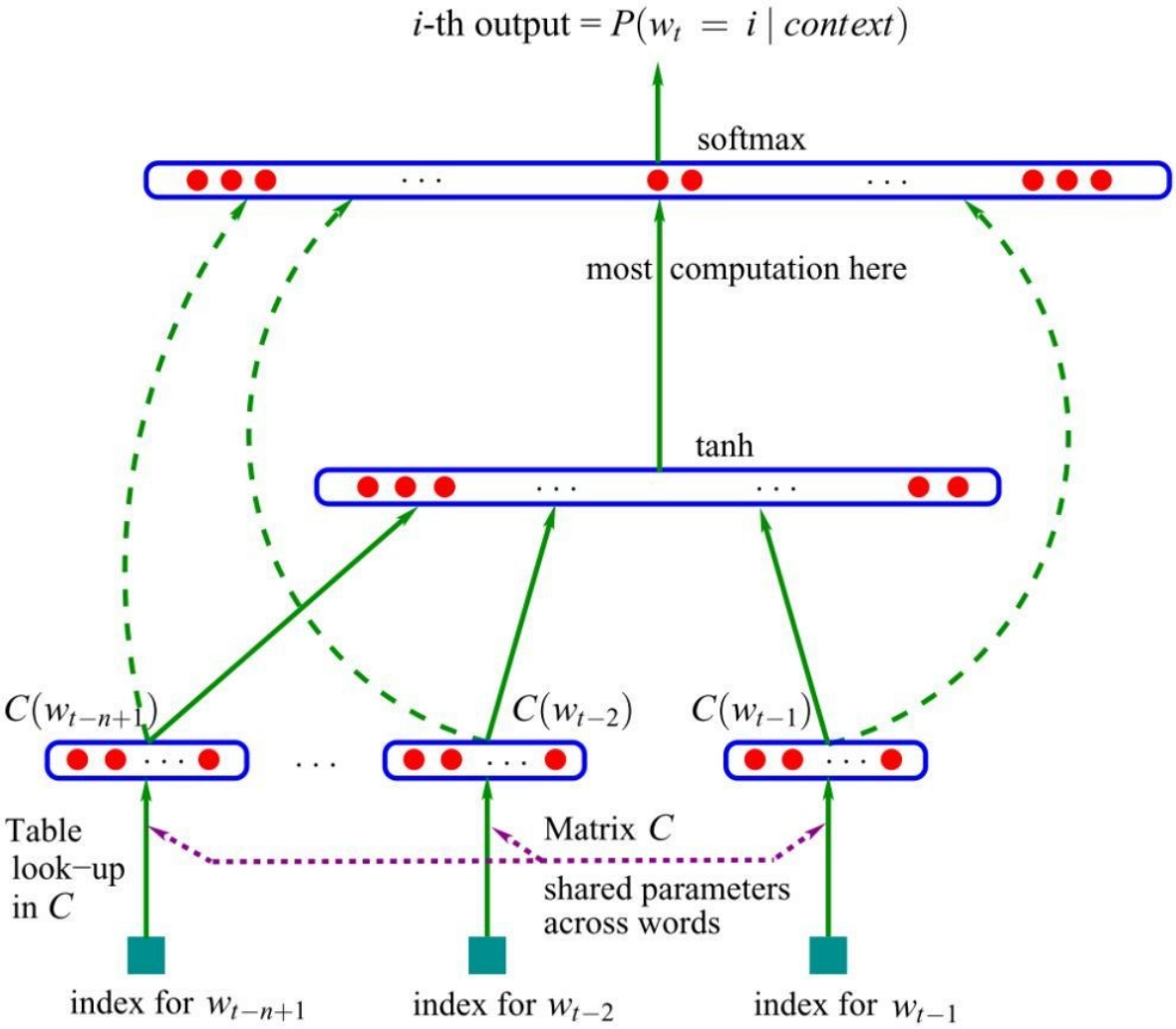


图 3 前馈神经网络语言模型(Bengio et al., 2001; 2003)

而现在构建语言模型的前馈神经网络，已被循环神经网络(RNNs)和长短期记忆神经网络(LSTMs)取代。虽然后来提出许多新模型在经典 LSTM 上进行了扩展，但它仍然是强有力的基础模型。甚至 Bengio 等人的经典前馈神经网络在某些设定下也和更复杂的模型效果相当，因为这些任务只需要考虑邻近的词语。理解这些语言模型究竟捕捉了哪些信息，也是当今一个活跃的研究领域。

语言模型的建立是一种**无监督学习(unsupervised learning)**，**Yann LeCun** 称之为**预测学习(predictive learning)**，是获得世界如何运作常识的先决条件。关于语言模型最引人注目的是，尽管它很简单，但却与后文许多核心进展息息相关。反过来，这也意味着 NLP 领域许多重要进展都可以简化为某种形式的语言模型构建。但要实现对自然语言真正意义上的理解，仅仅从原始文本中进行学习是不够的，我们需要新的方法和模型。

2.5 里程碑五：2008多任务学习(Multi-task learning)

多任务学习是在多个任务下训练的模型之间共享参数的方法，在神经网络中通过捆绑不同层的权重轻松实现。多任务学习思想 1993 年 Rich Caruana 首次提出，并应用于道路追踪和肺炎预测。多任务学习鼓励模型学习对多个任务有效的表征描述。这对于学习一般的、低级的描述形式、集中模型的注意力或在训练数据有限的环境中特别有用。

多任务学习 2008 年被 Collobert 和 Weston 等人首次在自然语言处理领域应用于神经网络。在他们的模型中，词嵌入矩阵被两个在不同任务下训练的模型共享，如图 4 所示。

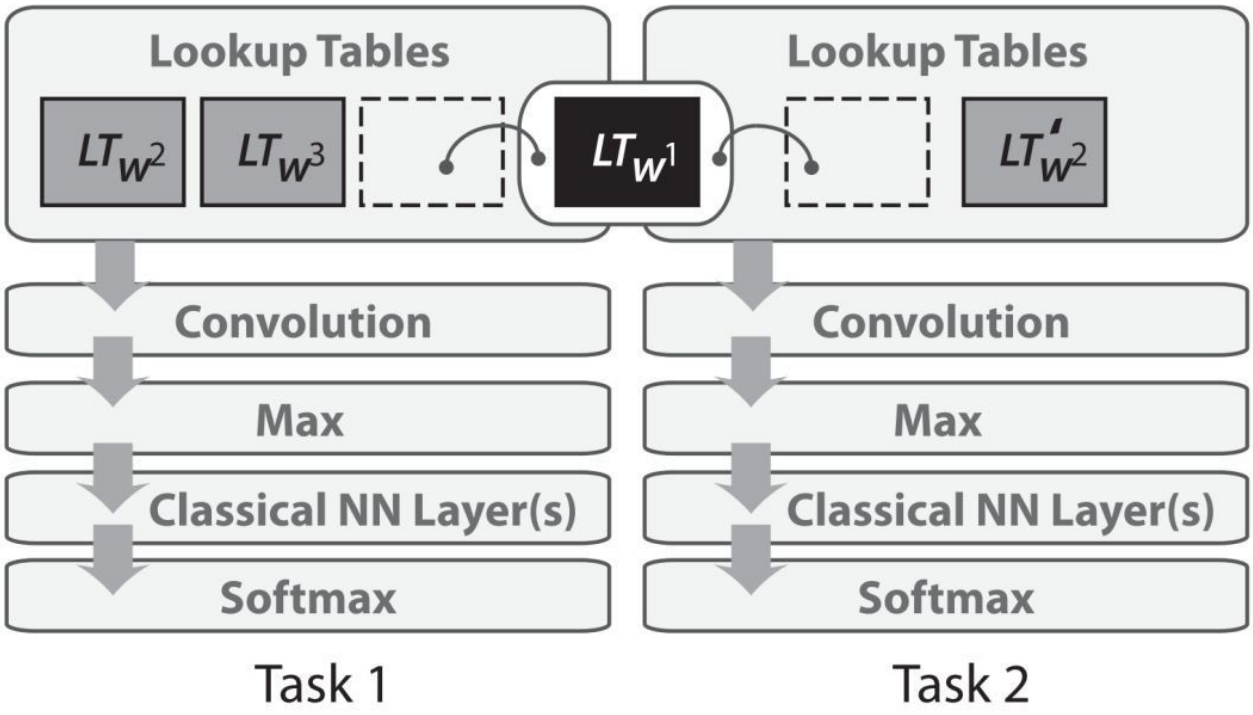


图 4 词嵌入矩阵共享(Collobert & Weston, 2008; Collobert et al., 2011)

共享的词嵌入矩阵使模型可以相互协作，共享矩阵中的低层级信息，而词嵌入矩阵往往构成了模型中需要训练的绝大部分参数。Collobert 和 Weston 发表于 2008 年的论文，影响远远超过了它在多任务学习中的应用。它开创的诸如预训练词嵌入和使用卷积神经网络处理文本的方法，在接下来的几年被广泛应用。他们也因此获得 **2018 年机器学习国际会议(ICML)的 test-of-time 奖**。

如今，多任务学习在自然语言处理领域广泛使用，而利用现有或“人工”任务已经成为 NLP 指令库中的一个有用工具。虽然参数的共享是预先定义好的，但在优化的过程中却可以学习不同的共享模式。当模型越来越多地在多个任务上进行测评以评估其泛化能力时，多任务学习就变得愈加重要，近年来也涌现出更多针对多任务学习的评估基准。

2.6 里程碑六：2013词嵌入

稀疏向量对文本进行表示的词袋模型，在自然语言处理领域有很长历史。而用**稠密的向量对词语进行描述，也就是词嵌入，则在 2001 年首次出现。**2013 年 **Mikolov** 等人工作主要创新之处在于，通过去除隐藏层和近似计算目标使词嵌入模型的训练更为高效。尽管这些改变本质上十分简单，但它们与高效的 word2vec(word to vector 用来产生词向量的相关模型)组合在一起，使得大规模的词嵌入模型训练成为可能。

Word2vec 有两种不同的实现方法：**CBOW**(continuous bag-of-words)和 **skip-gram**。它们在预测目标上有所不同：**一个是根据周围的词语预测中心词语，另一个则恰恰相反**。如图 5 所示。

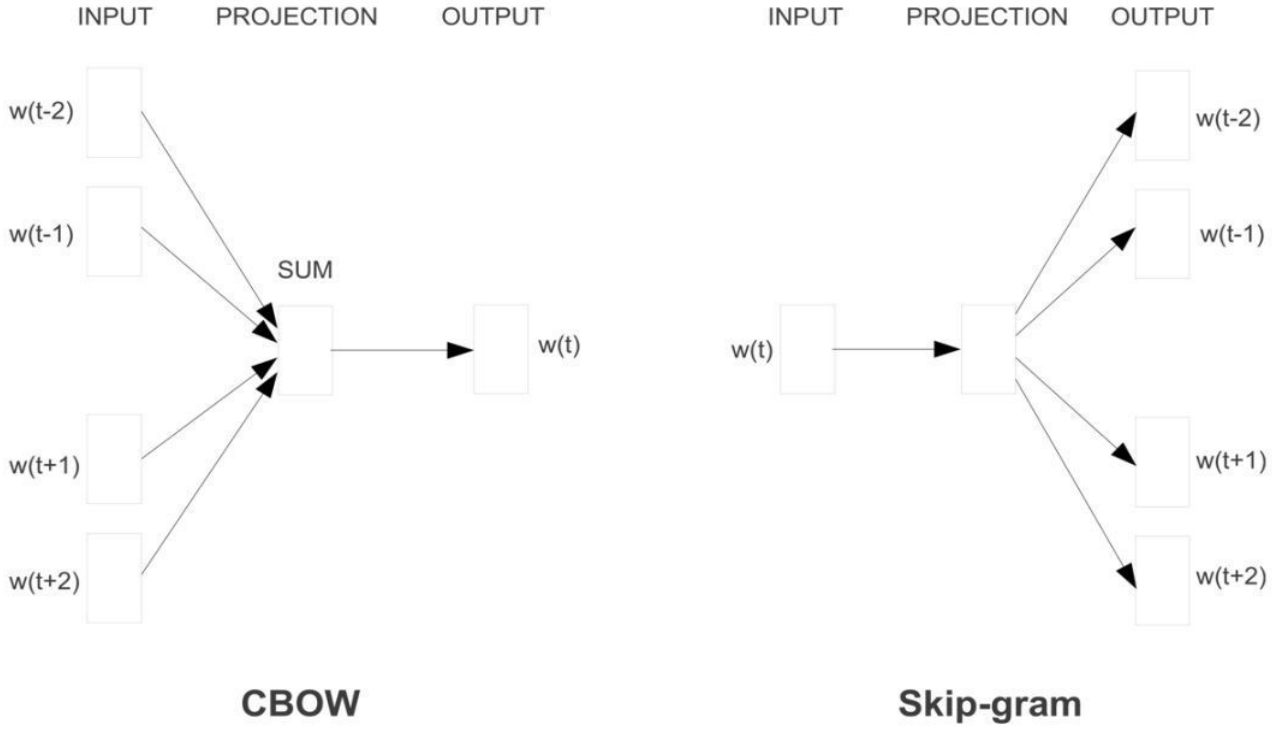


图 5 CBOW 和 skip-gram 架构(Mikolov et al., 2013a; 2013b)

虽然这些嵌入与使用前馈神经网络学习的嵌入在概念上没有区别，但是在一个非常大语料库上的训练使它们能够获取诸如性别、动词时态和国际事务等单词之间的特定关系。如下图 4 所示。

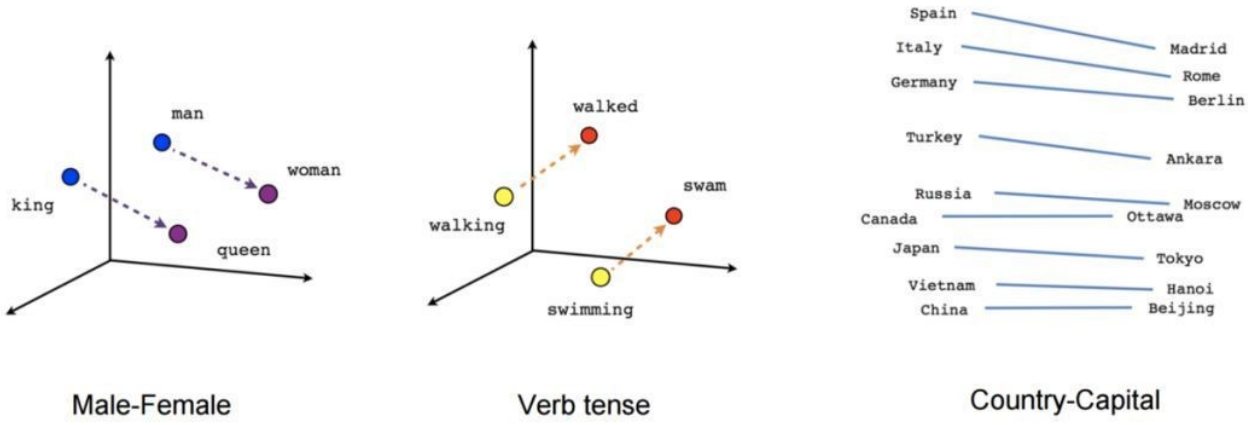


图 4 word2vec 捕获的联系(Mikolov et al., 2013a; 2013b)

这些关系和它们背后的意义激起了人们对词嵌入的兴趣，许多研究都在关注这些线性关系的来源。然而，**使词嵌入成为目前自然语言处理领域中流砥柱的，是将预训练的词嵌入矩阵用于初始化可以提高大量下游任务性能的事实**。

虽然 word2vec 捕捉到的关系具有直观且几乎不可思议的特性，但后来的研究表明，word2vec 本身并没有什么特殊之处：词嵌入也可以通过矩阵分解来学习，经过适当的调试，经典的矩阵分解方法 SVD 和 LSA 都可以获得相似的结果。从那时起，大量的工作开始探索词嵌入的不同方面。尽管有很多发展，word2vec 仍是目前应用最为广泛的选择。**Word2vec 应用范围也超出了词语级别**：带有负采样的 skip-gram——一个基于上下文学习词嵌入的方便目标，已经被用于学习句子的表征。它甚至超越了自然语言处理的范围，被应用于网络和生物序列等领域。

一个激动人心的研究方向是**在同一空间中构建不同语言的词嵌入模型，以达到(零样本)跨语言转换的目的**。通过无监督学习构建这样的映射变得越来越有希望(至少对于相似的语言来说)，这也为语料资源较少的语言和无监督机器翻译的应用程序创造可能。

2.7 里程碑七：2013RNN/CNN用于NLP的神经网络

2013 和 2014 年是自然语言处理领域神经网络时代的开始。其中三种类型的神经网络应用最为广泛：**循环神经网络**(recurrent neural networks)、**卷积神经网络**(convolutional neural networks)和**结构递归神经网络**(recursive neural networks)。

循环神经网络是 NLP 领域处理动态输入序列最自然的选择。Vanilla 循环神经网络很快被经典的**长短期记忆网络**(long-shortterm memory networks，LSTM)代替，该模型能更好地解决梯度消失和梯度爆炸问题。在 2013 年之前，人们仍认为循环神经网络很难训练，直到 Ilya Sutskever 博士的论文改变了循环神经网络这一名声。双向的长短期记忆记忆网络通常被用于同时处理出现在左侧和右侧的文本内容。LSTM 结构如图 7 所示。

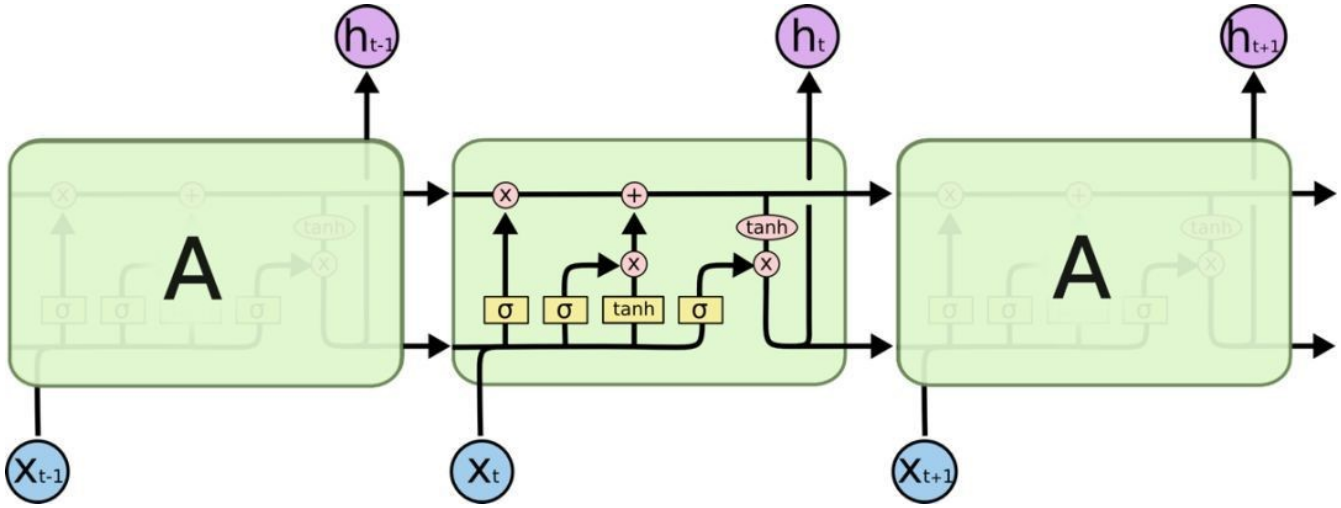


图 7 LSTM 网络(来源：ChrisOlah)

应用于文本的卷积神经网络只在两个维度上进行操作，卷积层只需要在时序维度上移动即可。图 8 展示了应用于自然语言处理的卷积神经网络的典型结构。

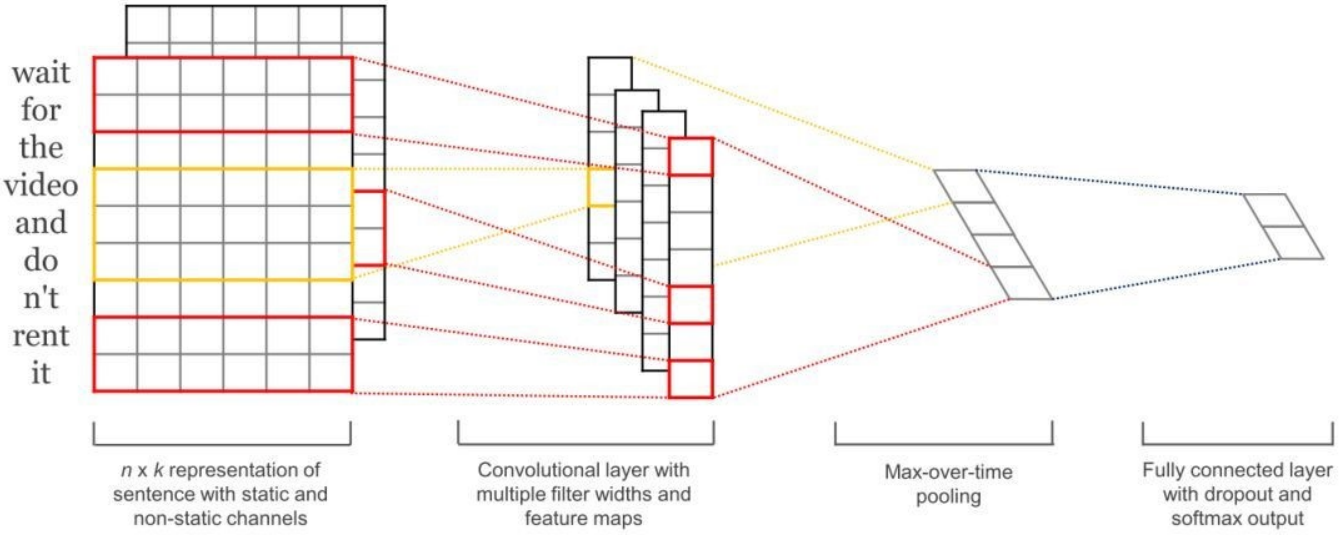


图 8 卷积神经网络(Kim,2014)

与循环神经网络相比，卷积神经网络的一个优点是具有更好的并行性。因为卷积操作中每个时间步的状态只依赖于局部上下文，而不是循环神经网络中那样依赖于所有过去的状态。卷积神经网络可以使用更大的卷积层涵盖更广泛的上下文内容。卷积神经网络也可以和长短期记忆网络进行组合和堆叠，还可以用来加速长短期记忆网络的训练。

循环神经网络和卷积神经网络都将语言视为一个序列。但从语言学的角度来看，语言是具有层级结构的：词语组成高阶的短语和小句，它们本身可以根据一定的产生规则递归地组合。这激发了利用结构递归神经网络，以**树形结构取代序列**来表示语言的想法，如图 9 所示。

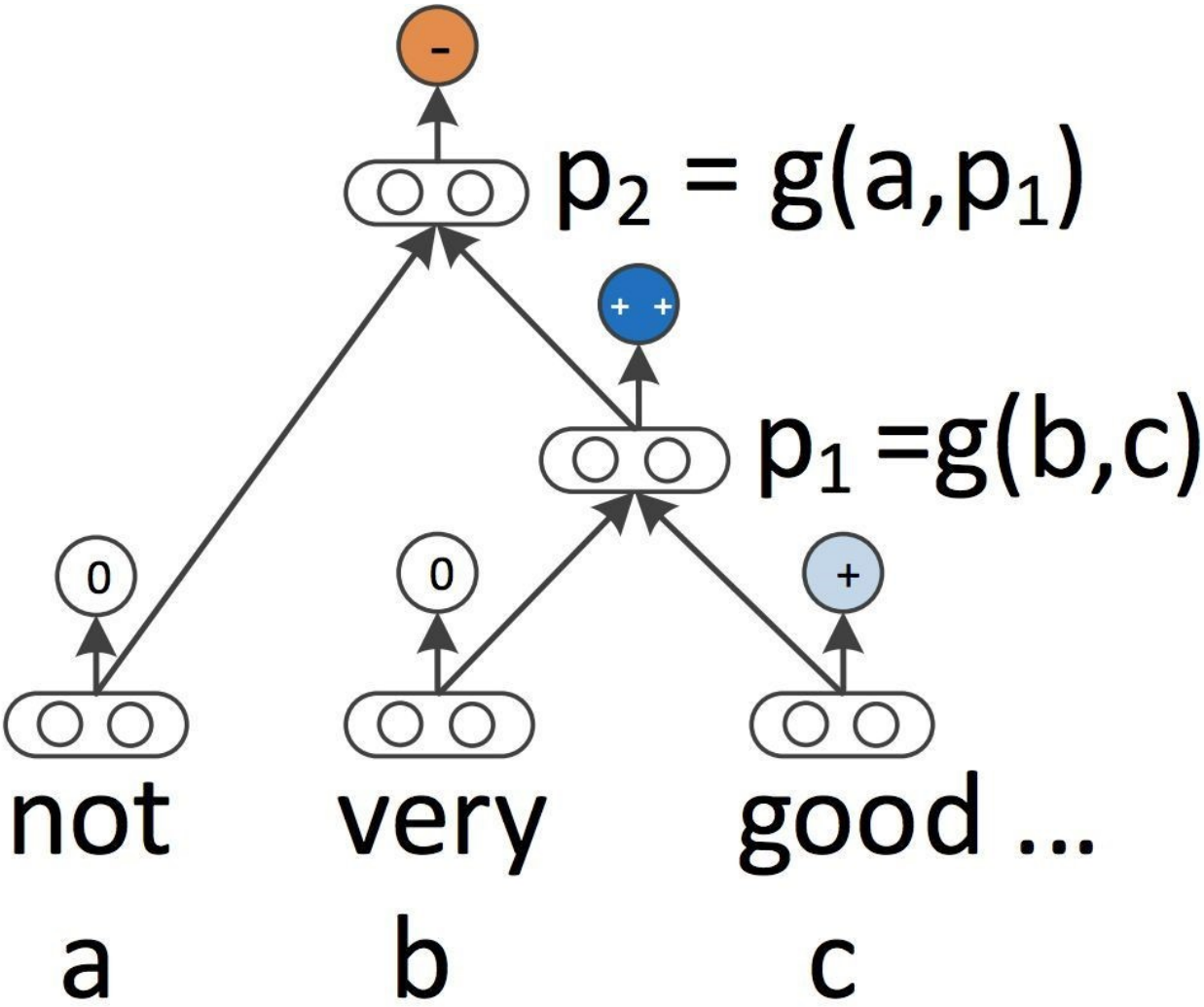


图 9 结构递归神经网络(Socher et al., 2013)

结构递归神经网络**自下而上构建序列**的表示，与从左至右或从右至左对序列进行处理的循环神经网络形成鲜明的对比。树中的每个节点是通过子节点的表征计算得到的。一个树也可以视为在循环神经网络上施加不同的处理顺序，所以长短期记忆网络则可以很容易地被扩展为一棵树。

不只是循环神经网络和长短期记忆网络可以扩展到使用层次结构，词嵌入也可以在语法语境中学习，语言模型可以基于句法堆栈生成词汇，图形卷积神经网络可以树状结构运行。

2.8 里程碑八：2014序列到序列模型(Sequence-to-sequencemodels)

2014 年，**Sutskever** 等人提出序列到序列学习，即使用神经网络将一个序列映射到另一个序列的一般化框架。在这个框架中，**一个作为编码器的神经网络**对句子符号进行处理，并将其压缩成向量表示；然后，**一个作为解码器的神经网络**根据编码器的状态逐个预测输出符号，并将前一个预测得到的输出符号作为预测下一个输出符号的输入。如图 10 所示。

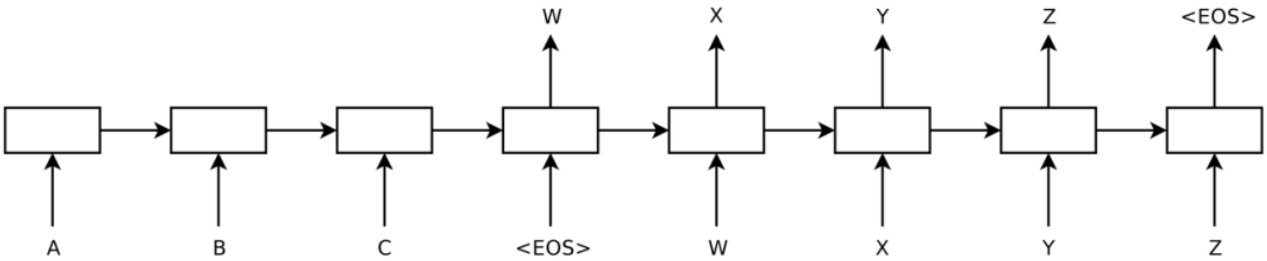


图 10 序列到序列模型(Sutskever et al., 2014)

机器翻译是这一框架的杀手级应用。2016 年，谷歌宣布他们将用神经机器翻译模型取代基于短语的整句机器翻译模型。谷歌大脑负责人 **Jeff Dean** 表示，这意味着**用 500 行神经网络模型代码取代 50 万行基于短语的机器翻译代码**。

由于其灵活性，该框架在自然语言生成任务上被广泛应用，其编码器和解码器分别由不同的模型来担任。更重要的是，解码器不仅可以适用于序列，在任意表示上均可以应用。比如**基于图片生成描述**(如图 11)、基于表格生成文本、根据源代码改变生成描述，以及众多其他应用。

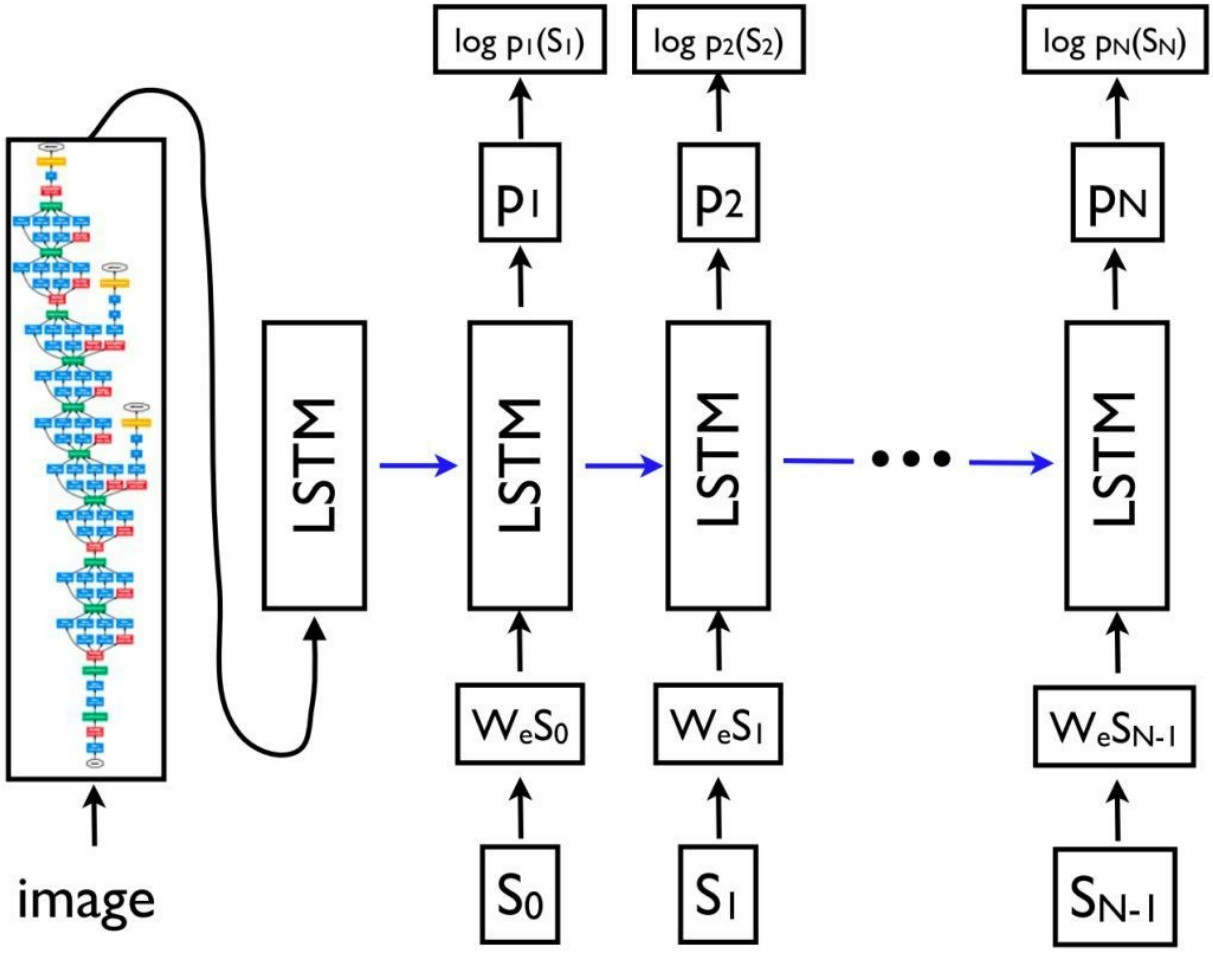


图 11 基于图像生成标题(Vinyalset al., 2015)

序列到序列的学习甚至可以应用到自然语言处理领域常见的结构化预测任务中，也就是输出具有特定的结构。为简单起见，**输出就像选区解析一样被线性化**(如图 12)。在给定足够多训练数据用于语法解析的情况下，神经网络已经被证明具有产生线性输出和识别命名实体的能力。

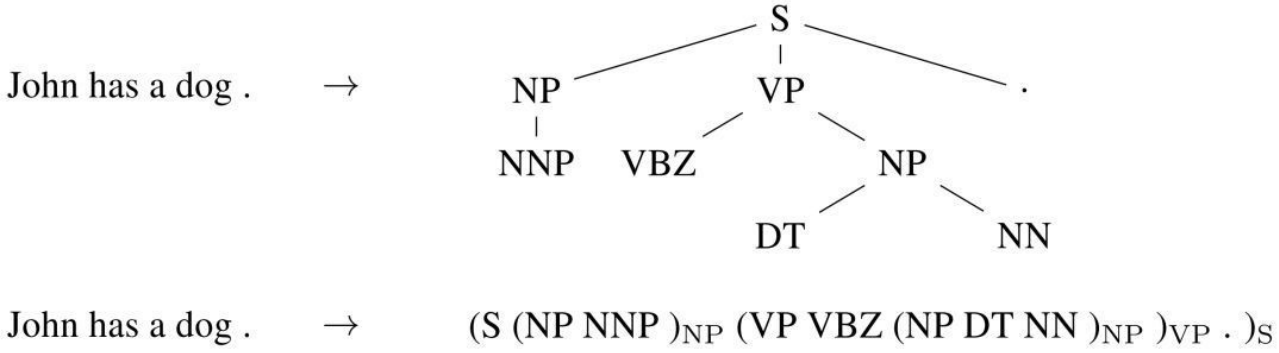


图 12 线性化选区解析树(Vinyalset al., 2015)

序列的编码器和解码器通常都是基于循环神经网络，但也可以使用其他模型。新的结构主要都从机器翻译的工作中诞生，它已经成了序列到序列模型的培养基。近期提出的模型有**深度长短期记忆网络、卷积编码器、Transformer**(一个基于自注意力机制的全新神经网络架构)以及**长短期记忆依赖网络和的 Transformer 结合体**等。

2.9 里程碑九：2015注意力机制和基于记忆的神经网络

注意力机制是神经网络机器翻译(NMT)的核心创新之一，也是使神经网络机器翻译优于经典的基于短语的机器翻译的关键。序列到序列学习的主要瓶颈是，需要将源序列的全部内容压缩为固定大小的向量。注意力机制通过让解码器回顾源序列的隐藏状态，以此为解码器提供加权平均值的输入来缓解这一问题，如图 13 所示。

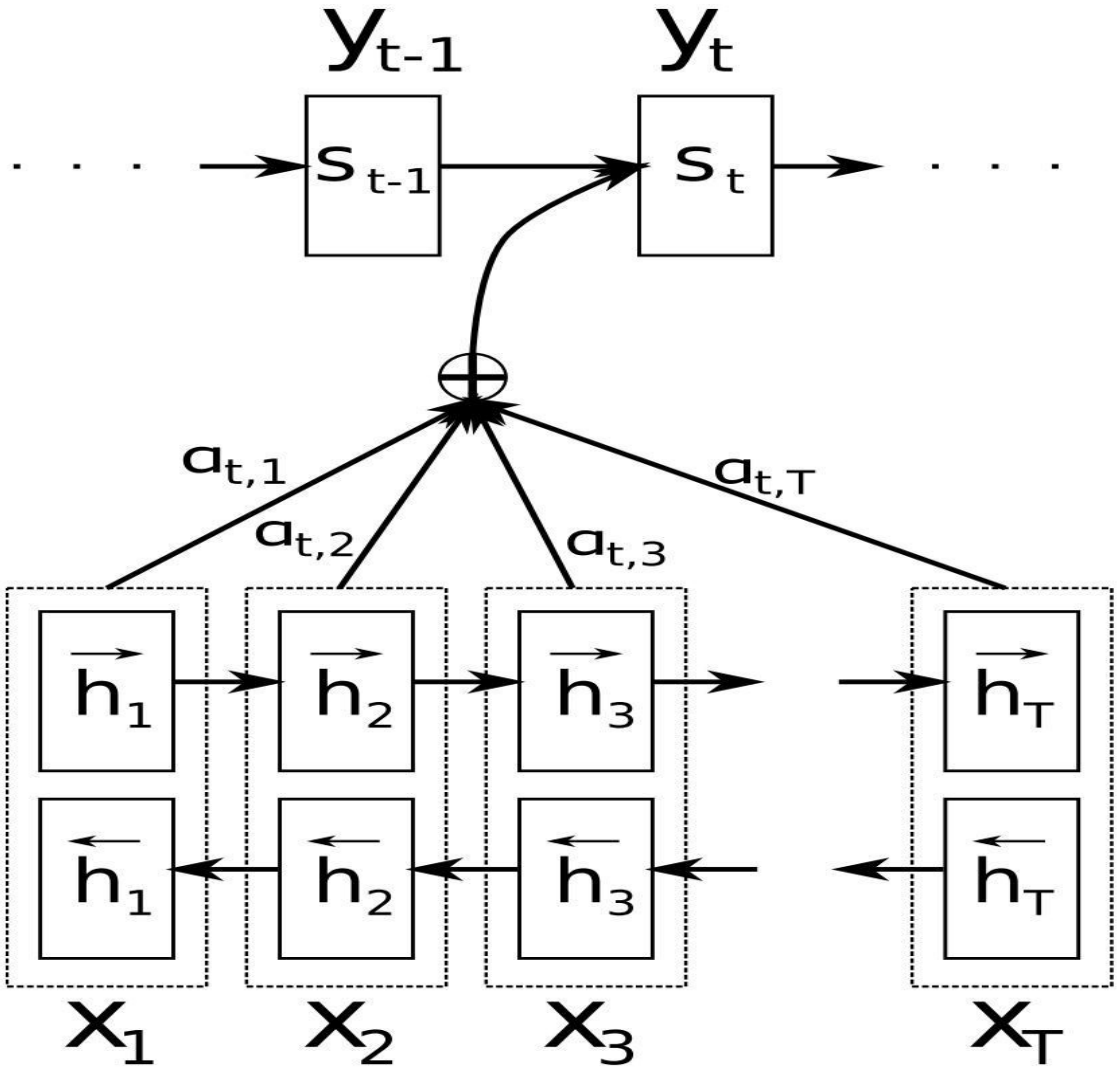


图 13 注意力机制(Bahdanau et al., 2015)

之后，各种形式的注意力机制涌现而出。注意力机制被广泛接受，在各种需要根据输入的特定部分做出决策的任务上都有潜在的应用。它已经**被应用于句法分析、阅读理解、单样本学习等任务中**。它的输入甚至不需要是一个序列，而可以包含其他表示，比如图像的描述(图 14)。

注意力机制一个有用的附带作用是它通过注意力权重来检测输入的哪一部分与特定的输出相关，从而提供了一种罕见的虽然还是比较浅层次的，对模型内部运作机制的窥探。

图 14 图像描述模型中的视觉注意力机制指示在生成”飞盘”时所关注的内容(Xu etal., 2015)

注意力机制不仅仅局限于输入序列。自注意力机制可用来观察句子或文档中周围的单词，获得包含更多上下文信息的词语表示。多层的自注意力机制是神经机器翻译前沿模型 Transformer 的核心。

注意力机制可以视为模糊记忆的一种形式，其记忆的内容包括模型之前的隐藏状态，由模型选择**从记忆中检索哪些内容。**与此同时，更多具有明确记忆单元的模型被提出。他们有很多不同的变化形式，比如神经图灵机(Neural Turing Machines)、记忆网络(Memory Network)、端到端的记忆网络(End-to-end Memory Newtorks)、动态记忆网络(DynamicMemory Networks)、神经可微计算机(Neural Differentiable Computer)、循环实体网络(RecurrentEntity Network)。

记忆的存取通常与注意力机制相似，基于与当前状态且可以读取和写入。这些模型之间的差异体现在它们如何实现和利用存储模块。比如说，端到端的记忆网络对输入进行多次处理并更新内存，以实行多次推理。神经图灵机也有一个基于位置的寻址方式，使它们可以学习简单的计算机程序，比如排序。基于记忆的模型通常用于需要长时间保留信息的任务中，例如语言模型构建和阅读理解。记忆模块的概念非常通用，知识库和表格都可以作为记忆模块，记忆模块也可以基于输入的全部或部分内容进行填充。

2.10 里程碑十：2018预训练语言模型

预训练的词嵌入与上下文无关，仅用于初始化模型中的第一层。近几个月以来，许多有监督的任务被用来预训练神经网络。相比之下，语言模型只需要未标记的文本，因此其训练可以扩展到数十亿单词的语料、新的领域、新的语言。预训练的语言模型于 2015 年被首次提出，但直到最近它才被证明在大量不同类型的任务中均十分有效。语言模型嵌入可以作为目标模型中的特征，或者根据具体任务进行调整。如下图所示，语言模型嵌入为许多任务的效果带来了巨大的改进。

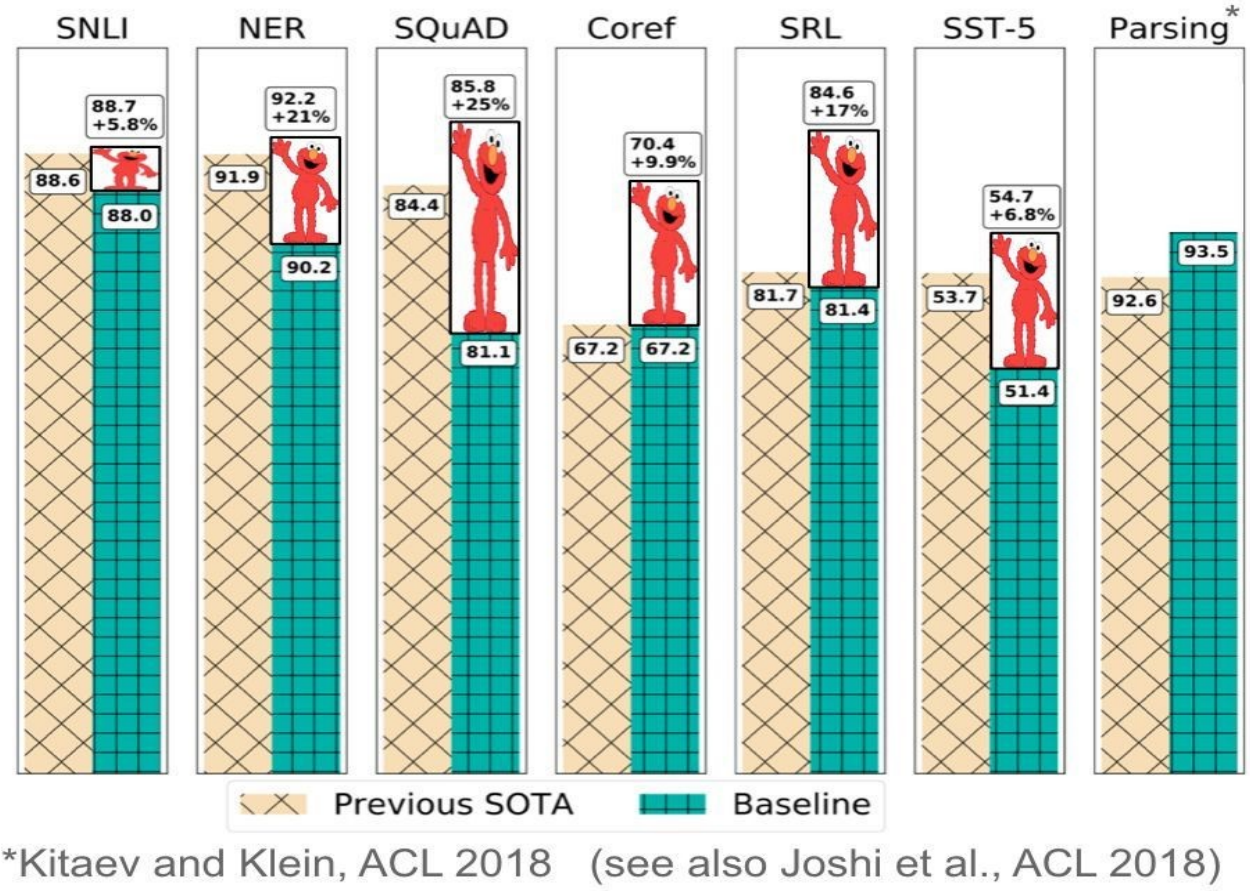


图 13 改进的语言模型嵌入(Peterset al., 2018)

使用预训练的语言模型可以在数据量十分少的情况下有效学习。由于语言模型的训练只需要无标签的数据，因此他们对于数据稀缺的低资源语言特别有利。2018 年 10 月，谷歌 AI 语言组发布 **BERT 语言模型预训练**，已被证明可有效改进许多自然语言处理任务(Dai and Le, 2015; Peters et al., 2017, 2018; Radford et al., 2018; Howard and Ruder, 2018)。这些任务包括句子级任务，如**自然语言推理 inference**(Bowman et al., 2015; Williams et al., 2018)和**释义 paraphrasing**(Dolan and Brockett, 2005)，旨在通过整体分析来预测句子之间的关系；以及词块级任务，如**命名实体识别**(Tjong Kim Sang and De Meulder, 2003)和 **SQuAD 问题回答**(Rajpurkar et al., 2016)，其中模型需要在词块级别生成细粒度输出。

近年七大技术里程碑小结

除了上述七大技术里程碑，一些其他进展虽不如上面提到的那样流行，但仍产生了广泛的影响。

基于字符的描述(Character-based representations)，在字符层级上使用卷积神经网络和长短期记忆网络，以获得一个基于字符的词语描述，目前已经相当常见了，特别是对于那些语言形态丰富的语种或那些形态信息十分重要、包含许多未知单词的任务。据目前所知，基于字符的描述最初用于序列标注，现在，基于字符的描述方法，减轻了必须以增加计算成本为代价建立固定词汇表的问题，并使完全基于字符的机器翻译的应用成为可能。

对抗学习(Adversarial learning)，在机器学习领域已经取得了广泛应用，在自然语言处理领域也被应用于不同的任务中。对抗样例的应用也日益广泛，他们不仅仅是探测模型弱点的工具，更能使模型更具鲁棒性(robust)。(虚拟的)对抗性训练，也就是最坏情况的扰动，和域对抗性损失(domain-adversariallosses)都是可以使模型更具鲁棒性的有效正则化方式。生成对抗网络(GANs)目前在自然语言生成任务上还不太有效，但在匹配分布上十分有用。

强化学习(Reinforcement learning)，在具有时间依赖性任务上证明有效，比如在训练期间选择数据和对话建模。在机器翻译和概括任务中，强化学习可以有效地直接优化“红色”和“蓝色”这样不可微的度量，不必去优化像交叉熵这样的代理损失函数。同样，逆向强化学习(inverse reinforcement learning)在类似视频故事描述这样的奖励机制非常复杂且难以具体化的任务中，也非常有用。

B 自然语言处理 NLP 知识结构(6990 字)

自然语言处理 NLP 知识结构

文|秦陇纪，数据简化 DataSimp20181013Sat

自然语言处理(计算机语言学、自然语言理解)涉及：字处理，词处理，语句处理，篇章处理词处理分词、词性标注、实体识别、词义消歧语句处理句法分析(Syntactic Analysis)、语义分析(Senmantic Analysis)等。其中，重点有：1.句法语义分析：分词，词性标记，命名实体识别。2.信息抽取3.文本挖掘：文本聚类，情感分析。基于统计。4.机器翻译：基于规则，基于统计，基于神经网络。5.信息检索6.问答系统7.对话系统建议...本文总结的自然语言处理历史、模型、知识体系结构内容，涉及NLP的语言理论、算法和工程实践各方面，内容繁杂。参考黄志洪老师自然语言处理课程、宗成庆老师《统计自然语言处理》，郑捷2017年电子工业出版社出版的图书《NLP汉语自然语言处理原理与实践》[29]，以及国外著名NLP书籍的英文资料、汉译版资料。[30]

一、NLP 知识结构概述

1)自然语言处理：利用计算机为工具，对书面实行或者口头形式进行各种各样的处理和加工的技术，是研究人与人交际中以及人与计算机交际中的演员问题的一门学科，是人工智能的主要内容。

2)自然语言处理是研究语言能力和语言应用的模型，建立计算机(算法)框架来实现这样的语言模型，并完善、评测、最终用于设计各种实用系统。

3)研究问题(主要)：

信息检索
机器翻译
文档分类
问答系统
信息过滤
自动文摘
信息抽取
文本挖掘
舆情分析
机器写作
语音识别

研究模式：*自然语言场景问题，数学算法，算法如何应用到解决这些问题，预料训练，相关实际应用*

自然语言的困难：

场景的困难：语言的多样性、多变性、歧义性
学习的困难：艰难的数学模型(hmm,crf,EM,深度学习等)
语料的困难：什么的语料？语料的作用？如何获取语料？

二、NLP 知识十大结构

2.1 形式语言与自动机

语言：按照一定规律构成的句子或者字符串的有限或者无限的集合。
描述语言的三种途径：
穷举法
文法(产生式系统)描述
自动机
自然语言不是人为设计而是自然进化的，形式语言比如：运算符号、化学分子式、编程语言
形式语言理论朱啊哟研究的是内部结构模式这类语言的纯粹的语法领域，从语言学而来，作为一种理解自然语言的句法规律，在计算机科学中，形式语言通常作为定义编程和语法结构的基础
形式语言与自动机基础知识：
集合论
图论
自动机的应用：
1，单词自动查错纠正
2，词性消歧(什么是词性？什么的词性标注？为什么需要标注？如何标注？)
形式语言的缺陷：
1、对于像汉语，英语这样的大型自然语言系统，难以构造精确的文法
2、不符合人类学习语言的习惯
3、有些句子语法正确，但在语义上却不可能，形式语言无法排出这些句子
4、解决方向：基于大量语料，采用统计学手段建立模型

2.2 语言模型

1)语言模型(重要)：通过语料计算某个句子出现的概率(概率表示)，常用的有 2-元模型，3-元模型
2)语言模型应用：
语音识别歧义消除例如，给定拼音串：ta shi yan yan jiu saun fa de
可能的汉字串：踏实烟酒算法的他是研究酸法的他是研究算法的，显然，最后一句才符合。
3)语言模型的启示：
1、开启自然语言处理的统计方法
2、统计方法的一般步骤：
收集大量语料
对语料进行统计分析，得出知识
针对场景建立算法模型
解释和应用结果
4)语言模型性能评价，包括评价目标，评价的难点，常用指标(交叉熵，困惑度)
5)数据平滑：
数据平滑的概念，为什么需要平滑
平滑的方法，加一法，加法平滑法，古德-图灵法，J-M 法，Katz 平滑法等
6)语言模型的缺陷：
语料来自不同的领域，而语言模型对文本类型、主题等十分敏感
n 与相邻的 n-1 个词相关，假设不是很成立。

2.3 概率图模型，生成模型与判别模型，贝叶斯网络，马尔科夫链与隐马尔科夫模型(HMM)

1)概率图模型概述(什么的概率图模型，参考清华大学教材《概率图模型》)
2)马尔科夫过程(定义，理解)
3)隐马尔科夫过程(定义，理解)

HMM 的三个基本问题(定义，解法，应用)

注：第一个问题，涉及最大似然估计法，第二个问题涉及 EM 算法，第三个问题涉及维特比算法，内容很多，要重点理解，(参考书李航《统计学习方法》，网上博客，笔者 github)

2.4 马尔科夫网，最大熵模型，条件随机场(CRF)

1)HMM 的三个基本问题的参数估计与计算

2)什么是熵

3)EM 算法(应用十分广泛，好好理解)

4)HMM 的应用

5)层次化马尔科夫模型与马尔科夫网络

提出原因，HMM 存在两个问题

6)最大熵马尔科夫模型

优点：与 HMM 相比，允许使用特征刻画观察序列，训练高效

缺点： 存在标记偏置问题

7)条件随机场及其应用(概念，模型过程，与 HMM 关系)

参数估计方法(GIS 算法，改进 IIS 算法)

CRF 基本问题：特征选取(特征模板)、概率计算、参数训练、解码(维特比)

应用场景：

词性标注类问题(现在一般用 RNN+CRF)

中文分词(发展过程，经典算法，了解开源工具 jieba 分词)

中文人名，地名识别

8)CRF++

2.5 命名实体 识别，词性标注，内容挖掘、语义分析与篇章分析(大量用到前面的算法)

1)命名实体识别问题

相关概率，定义

相关任务类型

方法(基于规程->基于大规模语料库)

2)未登录词的解决方法(搜索引擎，基于语料)

3)CRF 解决命名实体识别(NER)流程总结：

训练阶段：确定特征模板，不同场景(人名，地名等)所使用的特征模板不同，对现有语料进行分词，在分词结果基础上进行词性标注(可能手工)，NER 对应的标注问题是基于词的，然后训练 CRF 模型，得到对应权值参数值

识别过程：将待识别文档分词，然后送入 CRF 模型进行识别计算(维特比算法)，得到标注序列，然后根据标注划分出命名实体

4)词性标注(理解含义，意义)及其一致性检查方法(位置属性向量，词性标注序列向量，聚类或者分类算法)

2.6 句法分析

1)句法分析理解以及意义

1、句法结构分析

完全句法分析

浅层分析(这里有很多方法。。。)

2、 依存关系分析

2)句法分析方法

1、基于规则的句法结构分析

2、基于统计的语法结构分析

2.7 文本分类，情感分析

1)文本分类，文本排重

文本分类：在预定义的分类体系下，根据文本的特征，将给定的文本与一个或者多个类别相关联

典型应用：垃圾邮件判定，网页自动分类

2)文本表示，特征选取与权重计算，词向量

文本特征选择常用方法：

1、基于本文频率的特征提取法

2、信息增量法

3、X2(卡方)统计量

4、互信息法

3)分类器设计

SVM，贝叶斯，决策树等

4)分类器性能评测

1、召回率

2、正确率

3、F1 值

5)主题模型(LDA)与 PLSA

LDA 模型十分强大，基于贝叶斯改进了 PLSA，可以提取出本章的主题词和关键词，建模过程复杂，难以理解。

6)情感分析

借助计算机帮助用户快速获取，整理和分析相关评论信息，对带有感情色彩的主观文本进行分析，处理和归纳例如，评论自动分析，水军识别。

某种意义上看，情感分析也是一种特殊的分类问题

7)应用案例

2.8 信息检索，搜索引擎及其原理

1)信息检索起源于图书馆资料查询检索，引入计算机技术后，从单纯的文本查询扩展到包含图片，音视频等多媒体信息检索，检索对象由数据库扩展到互联网。

1、点对点检索

2、精确匹配模型与相关匹配模型

3、检索系统关键技术：标引，相关度计算

2)常见模型：布尔模型，向量空间模型，概率模型

3)常用技术：倒排索引，隐语义分析(LDA 等)

4)评测指标

2.9 自动文摘与信息抽取，机器翻译，问答系统

1)统计机器翻译的的思路，过程，难点，以及解决

2)问答系统

基本组成：问题分析，信息检索，答案抽取

类型：基于问题-答案， 基于自由文本

典型的解决思路

3)自动文摘的意义，常用方法

4)信息抽取模型(LDA 等)

2.10 深度学习在自然语言中的应用

1)单词表示，比如词向量的训练(wordvoc)

2)自动写文本

写新闻等

3)机器翻译

4)基于 CNN、RNN 的文本分类

5)深度学习与 CRF 结合用于词性标注

.....

更多深度学习内容，可参考我之前的文章。

自然语言处理(NLP)入门

本文简要介绍Python自然语言处理(NLP)，使用Python的NLTK库。NLTK是Python的自然语言处理工具包，在NLP领域中，最常使用的一个Python库。什么是NLP？简单来说，自然语言...

三、中文 NLP 知识目录

选自郑捷2017年电子工业出版社出版的图书《NLP汉语自然语言处理原理与实践》[29]。

C 自然语言处理 NLP 国内研究方向机构导师(1111 字)

自然语言处理 NLP 国内研究方向机构导师

文|中文信息协会《中文信息处理发展报告 2016》，数据简化 DataSimp20181021Sun

文字语言 VS 数字信息

数字、文字和自然语言一样，都是信息的载体，他们之间原本有着天然的联系。语言和数学的产生都是为了交流，从**文字、数字和语言**的发展历史，可以了解到语言、文字和数字有着内在的联系。**自然语言处理NLP**主要涉及三种文本，自由文本、结构化文本、半结构化文本。**自然语言理解Natural Language Understanding(NLU)**，实现人机间自然语言通信，意味着要使计算机既能理解自然语言文本的意义，也能以自然语言文本表达给定的意图、思想等。**自然语言生成NLG**，是人工或机器生成语言。斯坦福自然语言处理NLP工具资料收集、斯坦福分词、Stanford中文实体识别，最早做自然语言处理的网址https://nlp.stanford.edu/software/segmenter.shtml。哈尔滨工业大学智能技术与自然语言处理研究室(Intelligent Technology & Natural Language Processing Lab, ITNLPLab)是国内较早从事自然语言处理和语言智能技术的研究室。除了新兴的文本数据简化领域：秦陇纪(数据简化技术中心筹)，自然语言处理Natural Language Processing领域主要包括**基础研究**和**应用研究**。

基础研究

词法与句法分析：**李正华、陈文亮、张民**(苏州大学)
语义分析：**周国栋、李军辉**(苏州大学)
篇章分析：**王厚峰、李素建**(北京大学)
语言认知模型：**王少楠，宗成庆**(中科院自动化研究所)
语言表示与深度学习：**黄萱菁、邱锡鹏**(复旦大学)
知识图谱与计算：**李涓子、侯磊**(清华大学)

应用研究

文本分类与聚类：**涂存超，刘知远**(清华大学)
信息抽取：**孙乐、韩先培**(中国科学院软件研究所)
情感分析：**黄民烈**(清华大学)
自动文摘：**万小军、姚金戈**(北京大学)
信息检索：**刘奕群、马少平**(清华大学)
信息推荐与过滤：**王斌**(中科院信工所)，**鲁骁**(国家计算机网络应急中心)
自动问答：**赵军、刘康，何世柱**(中科院自动化研究所)
机器翻译：**张家俊、宗成庆**(中科院自动化研究所)
社会媒体处理：**刘挺、丁效**(哈尔滨工业大学)
语音技术：说话人识别——**郑方**(清华大学)，**王仁宇**(江苏师范大学)
语音合成——**陶建华**(中科院自动化研究所)
语音识别——**王东**(清华大学)
文字识别：**刘成林**(中科院自动化研究所)
多模态信息处理：**陈晓鸥**(北京大学)
医疗健康信息处理：**陈清财、汤步洲**(哈尔滨工业大学)
少数民族语言信息处理：**吾守尔•斯拉木**(新疆大学)

-End-

参考文献(4747 字)

1. Chomskyan linguistics encourages the investigation of "corner cases" that stress the limits of its theoretical models (comparable to pathological phenomena in mathematics), typically created using thought experiments, rather than the systematic investigation of typical phenomena that occur in real-world data, as is the case in corpus linguistics. The creation and use of such corpora of real-world data is a fundamental part of machine-learning algorithms for natural language processing. In addition, theoretical underpinnings of Chomskyan linguistics such as the so-called "poverty of the stimulus" argument entail that general learning algorithms, as are typically used in machine learning, cannot be successful in language processing. As a result, the Chomskyan paradigm discouraged the application of such models to language processing.

2. Jelinek, Frederick (1976). "Continuous speech recognition by statistical methods". Proceedings of the IEEE 64(4):532–556. doi:10.1109/PROC.1976.10159.

3. Garside, R., Leech, G. and Sampson, G. (eds.). The Computational Analysis of English: A Corpus-Based Approach. London: L

ongman, 1989.
4. David M. W. Powers and Christopher C. R. Turk (1989). Machine Learning of Natural Language. Springer-Verlag. ISBN 978-0-387-19557-5.
5. Jan Aarts, Willen Meijs (eds.). Corpus Linguistics: Theory and Practice. Amsterdam: Rodopi, 1990.
6. Hudson, R. A. English Word Grammar. Cambridge, Mass.: Basil Blackwell, 1991.
7. 白拴虎：《汉语词性自动标注系统研究》，清华大学计算机科学与技术系硕士学位论文，1992.
8. Bates, M (1995). "Models of natural language understanding". Proceedings of the National Academy of Sciences of the United States of America. 92 (22): 9977–9982. doi:10.1073/pnas.92.22.9977. PMC 40721.
9. M. Collins and J. Brooks. Preposition phrase attachment through a backed-off model. In Proceedings of the 3rd Workshop of Very Large Coepora, Cambridge, MA, 1995.
10. 董振东、董强：知网．《语言文字应用》1997(3) .
11. 俞士汶等：《现代汉语语法信息词典详解》．北京：清华大学出版社，1998.
12. Christopher D. Manning and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. The MIT Press. ISBN 978-0-262-13360-9.
13. Hutchins, J. (2005). "The history of machine translation in a nutshell".[self-published source] .
14. Daniel Jurafsky and James H. Martin (2008). Speech and Language Processing, 2nd edition. Pearson Prentice Hall. ISBN 978-0-13-187321-6.
15. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5. Official html and pdf versions available without charge.
16. Steven Bird, Ewan Klein, and Edward Loper (2009). Natural Language Processing with Python. O'Reilly Media. ISBN 978-0-596-51649-9.
17. Implementing an online help desk system based on conversational agent Authors: Alisa Kongthon, Chatchawal Sangkeettrakarn, Sarawoot Kongyoung and Choochart Haruechaiyasak. Published by ACM 2009 Article, Bibliometrics Data Bibliometrics. Published in: Proceeding, MEDES '09 Proceedings of the International Conference on Management of Emergent Digital EcoSystems, ACM New York, NY, USA. ISBN 978-1-60558-829-2, doi:10.1145/1643823.1643908.
18. Goldberg, Yoav (2016). A Primer on Neural Network Models for Natural Language Processing. Journal of Artificial Intelligence Research 57 (2016) 345–420.
19. Mohamed Zakaria Kurdi (2016). Natural Language Processing and Computational Linguistics: speech, morphology, and syntax, Volume 1. ISTE-Wiley. ISBN 978-1848218482.
20. Mohamed Zakaria Kurdi (2017). Natural Language Processing and Computational Linguistics: semantics, discourse, and applications, Volume 2. ISTE-Wiley. ISBN 978-1848219212.
21. Wikipedia. Natural language processing. [EB/OL]；Wikipedia，https://en.wikipedia.org/wiki/Natural_language_processing，2018-10-17.
22. 黄昌宁，张小凤，微软亚洲研究院．自然语言处理技术的三个里程碑．[EB/OL]；CSDN，https://blog.csdn.net/nuoline/article/details/8610661，2013-02-25.
23. Sebastian Ruder. A Review of the Neural History of Natural Language Processing. [EB/OL]；aylie，http://blog.aylien.com/a-review-of-the-recent-history-of-natural-language-processing/，2018-10-01.
24. 编辑：维尼，责编：王新凯．15 年来，自然语言处理发展史上的 8 大里程碑．[EB/OL]；搜狐科技，http://www.sohu.com/a/260525664_354973，2018-10-20.
25. 秦陇纪，数据简化 DataSimp 社区．理解和使用自然语言处理之终极指南(Python 编码)(经典收藏版 12k 字，附数据简化筹员 2 月 17 日 Fri 新闻)．[EB/OL]；CSDN，来源：数据简化 DataSimp(微信公众号)，https://blog.csdn.net/qq_28260611/article/details/58320374，2017-02-27.
26. 小郭．自然语言处理(NLP)知识结构总结．[EB/OL]；CSDN，https://blog.csdn.net/weixin_42137700/article/details/81983608，2018-08-23.
27. meihao5．自然语言处理(NLP)知识结构总结．[EB/OL]；CSDN，https://blog.csdn.net/meihao5/article/details/79592667，2018-03-17.
28. 创建词条：五巷 7 号(2017-11-18 10:48)，最近更新：小爱_四季私语(2018-05-15)．NLP 汉语自然语言处理原理与实践．[EB/OL]；百度百科，https://baike.baidu.com/item/NLP%E6%B1%89%E8%AF%AD%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86%E5%8E%9F%E7%90%86%E4%B8%8E%E5%AE%9E%E8%B7%B5/22211226，2018-05-15.
29. 郑捷．NLP 汉语自然语言处理原理与实践[C]；ISBN：9787121307652．千字数：816，页数：544，开本：16 开，出版时间：2017-01.
30. 中文信息协会．中文信息处理发展报告 2016[C]；国内关于自然语言处理的研究方向细分．[EB/OL]；https://blog.csdn.net/yezian d01/article/details/80525672，2016.
x.秦陇纪．数据简化社区 Python 官网 Web 框架概述；数据简化社区 2018 年全球数据库总结及 18 种主流数据库介绍；数据科学与大数据技术专业概论；人工智能研究现状及教育应用；信息社会的数据资源概论；纯文本数据溯源与简化之神经网络训练；大数据简化之技术体系．[EB/OL]；数据简化 DataSimp(微信公众号)，http://www.datasimp.org，2017-06-06.

Appx(845 字).数据简化 DataSimp 社区简介

信息社会之**数据、信息、知识、理论**持续累积，远超**个人认知学习**的时间、精力和能力。应对大数据时代的数据爆炸、信息爆炸、知识爆炸，解决之道重在**数据简化(Data Simplification)**：简化减少知识、媒体、社交数据，使信息、数据、知识越来越简单，符合人与设备的负荷。**数据简化 2018 年会议**(DS2018)聚焦**数据简化技术(Data Simplification techniques)**：对各类数据从采集、处理、存储、阅读、分析、逻辑、形式等方 ose **做简化**，应用于信息及数据系统、知识工程、各类 Python Web 框架、物理空间表征、生物医学数据，数学统计、自然语言处理、机器学习技术、人工智能等领域。欢迎投稿**数据科学技术、简化实例相关论文**提交电子版(最好有 PDF 格式)。填写申请表加入**数据简化 DataSimp** 社区成员，应至少一篇数据智能、编程开发 IT 文章：①高质量原创或翻译美欧数据**科技论文**；②社区网站义工或完善 **S 圈型**黑白静态和三彩色动态社区 **LOGO 图标**。**论文投稿**、加入**数据简化社区**，详情访问 www.datasimp.org 社区网站，网站维护请投会员邮箱 DataSimp@163.com。请关注公众号“数据简化

DataSimp”留言，或加微信 QinlongGEcai(备注：姓名/单位-职务/学校-专业/手机号)，免费加入**投稿群**或**”科学 Sciences 学术文献”**读者微信群等。长按下图“识别图中二维码”关注三个公众号(搜名称也行，关注后底部菜单有文章分类页链接)：

数据技术公众号“**数据简化 DataSimp**”：

科普公众号“科学 Sciences”：

社会教育知识公众号“**知识简化**”：