自然语言处理的形式模型

Formal Models of Natural Language Processing

冯志伟 著

中国科学技术大学出版社

当代科学技术基础理论与前沿问题研究丛书

自然语言处理的形式模型

Formal Models of Natural Language Processing

中国科学技术大学出版社

总　　序

侯建国

（中国科学技术大学校长、中国科学院院士、第三世界科学院院士）

大学最重要的功能是向社会输送人才.大学对于一个国家、民族乃至世界的重要性和贡献度,很大程度上是通过毕业生在社会各领域所取得的成就来体现的.

中国科学技术大学建校只有短短的五十年,之所以迅速成为享有较高国际声誉的著名大学之一,主要就是因为她培养出了一大批德才兼备的优秀毕业生.他们志向高远、基础扎实、综合素质高、创新能力强,在国内外科技、经济、教育等领域做出了杰出的贡献,为中国科大赢得了“科技英才的摇篮”的美誉.

2008年9月,胡锦涛总书记为中国科大建校五十周年发来贺信,信中称赞说:半个世纪以来,中国科学技术大学依托中国科学院,按照全院办校、所系结合的方针,弘扬红专并进、理实交融的校风,努力推进教学和科研工作的改革创新,为党和国家培养了一大批科技人才,取得了一系列具有世界先进水平的原创性科技成果,为推动我国科教事业发展和社会主义现代化建设做出了重要贡献.

据统计,中国科大迄今已毕业的5万人中,已有42人当选中国科学院和中国工程院院士,是同期(自1963年以来)毕业生中当选院士数最多的高校之一.其中,本科毕业生中平均每1 000人就产生1名院士和700多名硕士、博士,比例位居全国高校之首.还有众多的中青年才俊成为我国科技、企业、教育等领域的领军人物和骨干.在历年评选的“中国青年五四奖章”获得者中,作为科技界、科技创新型企业界青年才俊代表,科大毕业生已连续多年榜上有名,获奖总人数位居全国高校前列.鲜为人知的是,有数千名优秀毕业生踏上国防战线,为科技强军做出了重要贡献,涌现出20多名科技将军和一大批国防科技中坚.

为反映中国科大五十年来人才培养成果，展示毕业生在科学研究中的最新进展，学校决定在建校五十周年之际，编辑出版《中国科学技术大学校友文库》，于2008年9月起陆续出书，校庆年内集中出版50种.该《文库》选题经过多轮严格的评审和论证，入选书稿学术水平高，已列为“十一五”国家重点图书出版规划.

入选作者中，有北京初创时期的毕业生，也有意气风发的少年班毕业生；有“两院”院士，也有IEEE Fellow；有海内外科研院所、大专院校的教授，也有金融、IT行业的英才；有默默奉献、矢志报国的科技将军，也有在国际前沿奋力拼搏的科研将才；有“文革”后留美学者中第一位担任美国大学系主任的青年教授，也有首批获得新中国博士学位的中年学者……在母校五十周年华诞之际，他们通过著书立说的独特方式，向母校献礼，其深情厚意，令人感佩!

近年来，学校组织了一系列关于中国科大办学成就、经验、理念和优良传统的总结与讨论.通过总结与讨论，我们更清醒地认识到，中国科大这所新中国亲手创办的新型理工科大学所肩负的历史使命和责任.我想，中国科大的创办与发展，首要的目标就是围绕国家战略需求，培养造就世界一流科学家和科技领军人才.五十年来，我们一直遵循这一目标定位，有效地探索了科教紧密结合、培养创新人才的成功之路，取得了令人瞩目的成就，也受到社会各界的广泛赞誉.

成绩属于过去，辉煌须待开创.在未来的发展中，我们依然要牢牢把握“育人是大学第一要务”的宗旨，在坚守优良传统的基础上，不断改革创新，提高教育教学质量，早日实现胡锦涛总书记对中国科大的期待：瞄准世界科技前沿，服务国家发展战略，创造性地做好教学和科研工作，努力办成世界一流的研究型大学，培养造就更多更好的创新人才，为夺取全面建设小康社会新胜利、开创中国特色社会主义事业新局面贡献更大力量.

是为序.

2008年9月

前　言

采用计算机技术来研究和处理自然语言是从20世纪40年代末50年代初开始的，五十多年来，这项研究取得了长足的进展，成为了当代计算机科学中一门重要的新兴学科——自然语言处理(Natural Language Processing,NLP)。在信息网络时代，自然语言处理引起了包括计算机专家和语言学家在内的越来越多学者的重视，成为了一门文科和理科紧密结合的典型交叉学科。

由于现实的自然语言极为复杂，不可能直接作为计算机的处理对象，为了使现实的自然语言成为可以由计算机直接处理的对象，在自然语言处理的各个应用领域中，我们都需要根据处理的要求，把自然语言处理抽象为一个“问题”(problem)，再把这个问题在语言学上加以“形式化”(formalism)，建立语言的“形式模型”(formal model)，使之能以一定的数学形式，严密而规整地表示出来，并且把这种严密而规整的数学形式表示为“算法”(algorithm)，建立自然语言处理的“计算模型”(computational model)，使之能够在计算机上实现。在自然语言处理中，算法取决于形式模型，形式模型是自然语言计算机处理的本质，而算法只不过是实现形式模型的手段而已。这种建立语言形式模型的研究是非常重要的，它应当属于自然语言处理的基础理论研究。

本书对自然语言处理中的各种理论和方法进行了系统的总结和梳理，首先讨论了自然语言处理的学科定位，接着介绍了语言计算的一些先驱研究，然后以主要的篇幅讨论自然语言处理中的各种形式模型，包括基于短语结构语法的形式模型、基于合一运算的形式模型、基于依存和配价的形式模型、基于格语法的形式模型、基于词汇主义的形式模型、语义自动处理的形式模型、系统功能语法、语用自动处理的形式模型、概率语法、Bayes公式与动态规划算法、N-元语法和数据平滑、隐马尔可夫模型(HMM)、统计机器翻译的形式模型，同时还讨论了自然语言处理系统的评测问题，最后从哲学的角度讨论了自然语言处理中的理性主义和经验主义，探索理性主义方法和经验主义方法相结合的途径。

早在20世纪50年代在北京大学求学时，我就对自然语言的数学模型研究产生了兴趣，毅然从理科转到文科，师从王力、岑麒祥、朱德熙等著名语言学家学习语言学，探讨语言研究中的数学方法，后来成为理论语言学的研究生，并试图从理论上探讨自然语言处理的形式模型。可惜不久就发生了“文化大革命”，我被迫到边疆当了一名中学物理教员。

1978年高考制度恢复，我考入了中国科学技术大学研究生院。入学之后，学校公派我到法国格勒诺布尔理科医科大学应用数学研究所（IMAG）自动翻译中心（CETA）留学，师从法国著名数学家、国际计算语言学委员会主席沃古瓦（B. Vauquois）教授，系统地学习计算机科学和数学的知识，并专门研究文理交叉的自然语言处理问题，把语言学、计算机科学和数学紧密地结合起来。

中国科学技术大学使我有机会重新回到自然语言处理的队伍，使我有可能为我毕生钟爱的这个学科尽自己绵薄之力，我永远无法忘怀中国科学技术大学对我的恩情。值中国科学技术大学成立50周年之际，谨以此书为母校50华诞献礼。

在母校成立50周年的时候，我自己从事自然语言处理也恰巧有50年的日子了。50年前，我还是一个19岁的不谙世事的青年，现在，我已经是年过70岁的白发苍苍的老人了，我们这一代人正在一天天地变老；然而，我们如痴如醉地钟爱着的自然语言处理却是一门新兴的学科，她还非常年轻，充满了青春的活力，尽管她还比较幼稚娇嫩，还不够成熟，但是她无疑地有着光辉的发展前景。我们个人的生命是有限的，而科学知识的探讨和研究却是无限的。我们个人渺小的生命与科学事业这棵常青的参天大树相比较，显得多么地微不足道，有如沧海之一粟。想到这些，怎不令我们感慨万千！“书山有路勤为径，学海无涯苦作舟”，我们应当勤苦地工作，把个人的有限的生命投入到无限的科学知识的探讨和研究中去，从而实现人生的价值。

在本书的写作过程中，我曾参考过国内外时贤著作多种，没有他们丰厚的研究成果，本书是不可能写出来的，在此，我对他们表示由衷的感谢，就不一一列名道谢了。

本书涉及语言学、计算机科学、数学等多个领域的知识，我自己水平有限，错误在所难免，敬请广大读者提出宝贵的意见。

冯志伟

2008年11月

目　次

第 1 章　自然语言处理的学科定位

采用计算机技术来研究和处理自然语言是 20 世纪 40 年代末 50 年代初才开始的，五十多年来，这项研究取得了长足的进展，成为当代计算机科学中一门重要的新兴学科——自然语言处理(Natural Language Processing，NLP)。在信息网络时代，自然语言处理引起了越来越多的学者的重视，成为一门“显学”，提出了各种不同的理论和方法。本书将对自然语言处理中的这些理论和方法进行系统的总结和梳理，重点研究自然语言处理这个新兴学科中的形式模型问题。

在工业革命时代，人类需要探索物质世界的奥秘，由于物质世界是由原子和各种基本粒子构成的，因此，研究原子和各种基本粒子的物理学成了非常重要的学科；在信息网络时代，由于信息网络主要是由语言构成的，因此，我们可以预见，在不久的将来，研究语言结构的自然语言处理必定也会成为像物理学一样非常重要的学科。物理学研究物质世界中各种物理运动的规律，而自然语言处理则研究信息网络世界中语言载体的规律。自然语言处理的重要性完全可以与物理学媲美，它们将成为未来科学世界中举足轻重的双璧。这是我们在直觉上的一种估计，我们坚信这样的估计将会成为活生生的现实。

在这样的情况下，如何对自然语言处理进行正确的学科定位，使我们认识到自然语言处理在整个学科体系中的位置，从而自觉地推动自然语言处理的发展，是一个至关重要的问题。

我们可以从自然语言处理的过程、自然语言处理的范围以及自然语言处理的历史三个角度来考察自然语言处理的学科定位问题。从自然语言处理的过程来考察它的学科定位，是从纵的角度来讨论；从自然语言处理的范围来考察它的学科定位，是从横的角度来讨论；纵横交错，我们对于自然语言处理的学科定位就可以在共时的平面上得到比较清晰的认识。最后，我们再从自然语言处理的历史来考察，也就是从发展的角度来讨论，这样，我们对于自然语言处理的学科定位就可以在历时的平面上得到比较清晰的认识。

1.1　从自然语言处理的过程来考察其学科定位

首先，我们从自然语言处理的过程，也就是从纵的角度来讨论这个问题。

我们认为，计算机对自然语言的研究和处理，一般应经过如下四个方面的过程：

● 把需要研究的问题在语言学上加以形式化，建立语言的形式化模型，使之能以一定的数学形式，严密而规整地表示出来，这个过程可以叫做“形式化”。

● 把这种严密而规整的数学形式表示为算法，这个过程可以叫做“算法化”。

● 根据算法编写计算机程序，使之在计算机上加以实现，建立各种实用的自然语言处理系统，这个过程可以叫做“程序化”。

● 对于所建立的自然语言处理系统进行评测，使之不断地改进质量和性能，以满足用户的要求，这个过程可以叫做“实用化”。

美国计算机科学家 Bill Manaris（马纳利斯）在 1999 年出版的《计算机进展》（*Advances in Computers*）第 47 卷的《从人-机交互的角度看自然语言处理》一文中曾经给自然语言处理提出了如下的定义：

“自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力（linguistic competence）和语言应用（linguistic performance）的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术。”这个定义的英文如下：“NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the result systems.”（Bill Manaris, Natural language processing: A human-computer interaction perspective, Advances in Computers, Volume 47,1999）

Bill Manaris 关于自然语言处理的这个定义，比较全面地表达了计算机对自然语言的研究和处理的上述四个方面的过程。我们认同这样的定义。

在 2001 年的美国电影《太空奥德赛》中（Stanley Kubrick 和 Arthur C. Charke 编，Screenplay of 2001: *A Space Odyssey*）机器人 HAL 和 Dave 进行了如下对话：

Dave Bownman: Open the pod bay doors, HAL.

HAL: I'm sorry Dave, I'am afraid I can't do that.

（Dave Bownman: HAL，请你打开太空舱的分离舱门。

HAL: 对不起，Dave，我不能这样做。）

HAL 实际上是一台名为“9000”的电子计算机，这台计算机具有 20 世纪最受人们认可的一些特征，HAL 实际上是一个具有高级的语言处理能力并且能够说英语和理解英语的智能机器人（artificial agent），在影片情节的关键时刻，HAL 甚至能够进行唇读（reading lip），上面就是电影中的角色 Dave 先生请求智能机器人 HAL 打开宇宙飞船的分离舱门（pod bay doors）与 HAL 之间的一段对话。HAL 的作者 Arthur C. Charke 曾经乐观地预言，到一定的时候，我们就可以制造出像 HAL 这样的智能机器人。但是，现在我们离这样的预言还有多远呢？为了让 HAL 具有与语言相关的能力，我们究竟还应该做些什么呢？

我们认为，像 HAL 这样的机器人至少应该通过语言与人类进行交流。其中包括通

过语音识别(speech recognition)和自然语言理解(natural language processing,当然包括唇读)来与人类沟通,通过自然语言生成(natural language generation)和语音合成(speech synthesis)来与人类交际。HAL 也应该能够做信息检索(information retrieval,发现它所需要的文本资源在哪里)和信息抽取(information extraction,从文本资源中抽取它所需要的信息),并且进行知识推理(reference,根据已知的事实推出结论)。

尽管这些问题现在还远远没有完全解决,HAL 需要的一些与语言相关的技术现在已经研制出来了,有一部分技术已经商品化。解决这样的问题以及其他类似的问题,是自然语言处理、计算语言学、语音识别与语音合成的主要研究内容。我们把它们统称为语音与语言的计算机处理(speech and language processing),或者简单地称为自然语言处理(natural language processing),因此,自然语言处理也同时包括了语音处理的内容。

像 HAL 这样有复杂的语言能力的智能机器人将要求非常广泛和深刻的语言知识。我们只要读一读前面在 HAL 和 Dave 之间进行的对话,我们就可以了解到这样的更加复杂的应用所需要的语言知识的范围和种类。

为了确定 Dave 讲什么,HAL 必须能够分析它所接收的声音信号,并且把 Dave 的这些信号复原成词的系列。与此相似,为了生成回答,HAL 必须把它的回答组织成词的系列,并且生成 Dave 能够识别的声音信号。要完成这两方面的任务,需要语音学(phonetics)和音系学(phonology)的知识,这样的知识可以帮助我们建立词如何在话语中发音的模型。

值得注意的是,HAL 还能够说出如像 I'm 和 can't 这样的缩约形式,HAL 必须把它们分别还原为 I am 和 can not,才能在它的词库中找到这些单词的对应物,从而明白这些缩约形式究竟代表什么样的语言成分。HAL 还要能够产生并且识别单词的这样或那样的变体(例如,识别 doors 是复数)。这些都要求 HAL 具有形态学方面的知识,这些知识能够反映关于上下文中词的形态和行为的有关信息。

除了处理一个一个的单词之外,HAL 还应该知道怎样分析 Dave 所提出的请求的结构。这样的分析能够使 HAL 确定,Dave 说的话是关于要 HAL 采取某种行动的一个请求,这样的请求不同于下面关于陈述客观世界的简单命题,也不同于下面关于 door 的问话,它们是 Dave 请求的不同变体:

HAL, the pod bay door is open.(HAL,分离舱的门是开着的。)

HAL, is the pod bay door open? (HAL,分离舱的门是开着的吗?)

此外,HAL 还必须使用类似的结构知识把一个个的单词组织成为符号串,构成它的回答。例如,HAL 必须知道,下面的单词序列对于 Dave 是没有意义的,尽管这个单词系列所包含的单词与它原来的回答中所包含的单词完全一样:

I'm I do, sorry that afraid Dave I'm can't.

这里所说的关于组词成句的知识,叫做句法(syntax)。

显而易见,如果只是知道 Dave 所说的话语的各个单词以及句法结构,并不能使 HAL 了解 Dave 提出的请求的实质。为了理解 Dave 的请求事实上是关于要求关闭 pod bay door(分离舱门)的一个命令,而不是讲关于当天中饭的菜单的事情,就要有复合词的语义的知识、词汇语义学(lexical semantics)的知识以及如何把这样的复合词组成更大的意义的知识,即关于组合语义学(compositional semantics)的知识。pod bay door 按

照字面逐词翻译是“豆荚-海湾-门”,但是它们组合成的意思却是“分离舱门”。这是关于科学技术术语(terminology)的知识。

另外,尽管智能机器人 HAL 的行为还不十分熟练,它也应该充分地懂得如何对 Dave 表示礼貌。例如,它不要简单地回答 No 或者 No, I won't open the door。HAL 首先用表示客气的话回答 I'm sorry 和 I'm afraid,然后委婉地说 I can't,而不是直截了当地说 I won't。这种礼貌和委婉语言的用法属于语用学(pragmatics)的研究领域。

最后,HAL 不是简单地无视 Dave 的请求,让门继续关着,而是对于 Dave 开始的请求,选择结构会话的方式来对待。HAL 在它给 Dave 的回答中,正确地使用单词 that 来简单地表示会话中话段之间的共同部分。正确地把这样的会话组织成结构,需要话语规约(discourse convention)的知识。

因此,我们认为,建立自然语言处理模型需要如下 9 个不同平面的知识:

- 声学和韵律学的知识:描述语言的节奏、语调和声调的规律,说明语音怎样形成音位。
- 音位学的知识:描述音位的结合规律,说明音位怎样形成语素。
- 形态学的知识:描述语素的结合规律,说明语素怎样形成单词。
- 词汇学的知识:描述词汇系统的规律,说明单词本身固有的语义特性和语法特性。
- 句法学的知识:描述单词(或词组)之间的结构规则,说明单词(或词组)怎样形成句子。
- 语义学的知识:描述句子中各个成分之间的语义关系,这样的语义关系是与情景无关的,说明怎样从构成句子的各个成分推导出整个句子的语义。
- 话语分析的知识:描述句子与句子之间的结构规律,说明怎样由句子形成话语或对话。
- 语用学的知识:描述与情景有关的情景语义,说明怎样推导出句子具有的与周围话语有关的各种含义。
- 外界世界的常识性知识:描述关于语言使用者和语言使用环境的一般性常识,例如,语言使用者的信念和目的,说明怎样推导出这样的信念和目的内在的结构。

当然,关于自然语言处理所涉及的知识平面还有不同的看法,不过,一般而言,大多数的自然语言处理研究人员都认为,这些语言学知识至少可以分为词汇学知识、句法学知识、语义学知识和语用学知识等平面。每一个平面传达信息的方式各不相同。例如,词汇学平面可能涉及具体的单词的构成成分(例如,语素)以及它们的屈折变化形式的知识;句法学平面可能涉及在具体的语言中单词或词组怎样结合成句子的知识;语义学平面可能涉及怎样给具体的单词或句子指派意义的知识;语用学平面可能涉及在对话中话语焦点的转移以及在给定的上下文中怎样解释句子含义的知识。

下面我们具体说明在自然语言处理中这些知识平面的一般情况。如果我们对计算机发一个口头的指令:“Delete file x”(“删除文件 x”),我们要通过自然语言处理系统让计算机理解这个指令的含义,并且执行这个指令,一般来说需要经过处理过程见图 1.1。

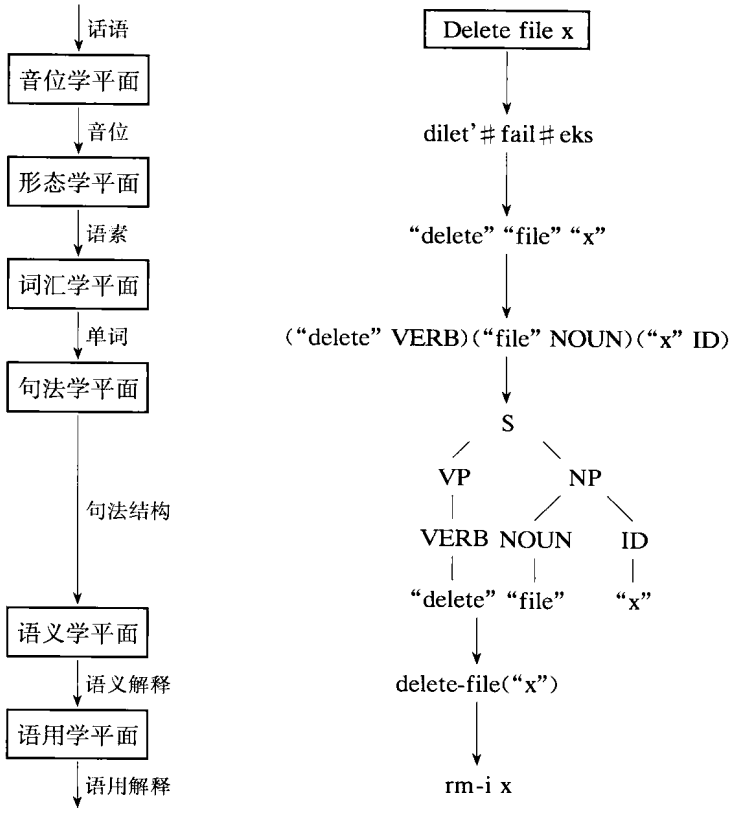

图 1.1 自然语言处理系统中的知识平面

从图 1.1 中可以看出，自然语言处理系统首先把指令"Delete file x"在音位学平面转化成音位系列"dilet'# fail # eks"，然后在形态学平面把这个音位系列转化为语素系列"delete" "file" "x"，接着在词汇学平面把这个语素系列转化为单词系列并标注相应的词性：("delete"VERB) ("file"NOUN) ("x"ID)，在句法学平面进行句法分析，得到这个单词系列的句法结构，用树形图表示，在语义学平面得到这个句法结构的语义解释：delete-file ("x")，在语用学平面得到这个指令的语用解释"rm-i x"，最后让计算机执行这个指令。

这个例子来自美国自然语言处理学者 Wilensky(威林斯基)为 UNIX 设计的一个语音理解界面，叫做 UNIX Consultant。这个语音理解界面使用了上述的第 1 至第 6 个平面的知识，得到口头指令"Delete file x"的语义解释：delete-file ("x")，然后，使用第 8 个平面的语用学知识把这个语义解释转化为计算机的指令语言"rm-i x"，让计算机执行这个指令，这样便可以使用口头指令来指挥计算机的运行了。

不同的自然语言处理系统需要的知识平面可能与 UNIX Consultant 不一样，根据实际应用的不同要求，很多自然语言处理系统只需要使用上述 9 个平面中的部分平面的知识就行了。例如，书面语言的机器翻译系统只需要第 3 至第 7 个平面的知识，个别的机器翻译系统还需要第 8 个平面的知识；语音识别系统只需要第 1 至第 5 个平面的知识。

上述 9 个平面的知识主要涉及的是语言学知识，所以我们认为自然语言处理原则上

是一个语言学问题。除了语言学之外，自然语言处理还涉及如下的知识领域：

- 计算机科学：给自然语言处理提供模型表示、算法设计和计算机实现的技术。
- 数学：给自然语言处理提供形式化的数学模型和形式化的数学方法。
- 心理学：给自然语言处理提供人类言语行为的心理模型和理论。
- 哲学：给自然语言处理提供关于人类的思维和语言的更深层次的理论。
- 统计学：给自然语言处理提供基于样本数据来预测统计事件的技术。
- 电子工程：给自然语言处理提供信息论的理论基础和语言信号处理技术。
- 生物学：给自然语言处理提供大脑中人类语言行为机制的理论。

因此，自然语言处理是一个多边缘的交叉学科。自然语言处理的研究，应该把这些学科的知识结合起来。每一个从事自然语言处理研究的人，都应该尽量使自己成为文理兼通、博学多识的人。

1.2　从自然语言处理的范围来考察其学科定位

上面，我们从自然语言处理的过程，也就是从纵的角度，考察了自然语言处理的学科定位。下面，我们换一个角度，从自然语言处理的范围，也就是从横的角度来考察自然语言处理的学科定位。

自然语言处理的范围涉及众多的部门，如语音的自动识别与合成、机器翻译、自然语言理解、人机对话、信息检索、文本分类、自动文摘，等等。我们认为，这些部门可以归纳为如下四个大的方向：

- 语言学方向：把自然语言处理作为语言学的分支来研究，它只研究语言及语言处理与计算相关的方面，而不管其在计算机上的具体实现。这个研究方向的最重要的研究领域是语法形式化理论和自然语言处理的数学理论。
- 数据处理方向：把自然语言处理作为开发语言研究相关程序以及语言数据处理的学科来研究。这一方向的研究早期有术语数据库的建设、各种机器可读的电子词典的开发，近年来随着大规模语料库的出现，这个方向的研究显得更加重要。
- 人工智能和认知科学方向：把自然语言处理作为在计算机上实现自然语言能力的学科来研究，探索自然语言理解的智能机制和认知机制。这一方向的研究与人工智能以及认知科学关系密切。
- 语言工程方向：把自然语言处理作为面向实践的、工程化的语言软件开发来研究。这一方向的研究一般称为“人类语言技术(Human Language Technique，HLT)”，或者称为“语言工程”(language engineering)。

最近，德国出版了一本叫做《计算语言学和语言技术》(*Computerlinguistik and Sprachtechnologie*)的专著，把目前自然语言处理的研究领域也分为四个方向

(Carstensen 2004),与我们的分法大致相同。

这四个方向的概括,大致涵盖当今自然语言处理研究的内容,更加细致地说,自然语言处理可以进一步细分为如下 13 个方面的内容:

1. 口语输入(spoken language input)

- 语音识别(speech recognition)。
- 信号表示(语音信号分析)(signal representation (voice signal analysis))。
- 鲁棒的语音识别(robust speech recognition)。
- 语音识别中的隐马尔可夫模型方法(HMM (Hidden Markov Model) methods in speech recognition)。
- 语言表示理论(语言模型)(language representation (language model))。
- 说话人识别(speaker recognition)。
- 口语理解(spoken language understanding)。

2. 书面语输入(written language input)

- 文献格式识别(document image (format) analysis)。
- 光学字符识别:印刷体识别(OCR (Optical Character Recognition): print)。
- 光学字符识别:手写体识别(OCR: handwriting)。
- 手写界面(例如,用笔输入的计算机)(handwriting as computer interface (e.g. pen computer))。
- 手写文字分析(例如,签名验证)(handwriting analysis (e.g. signature verification))。

3. 语言分析和理解(language analysis and understanding)

- 小于句子单位的处理(形态分析,形态排歧)(sub-sentential processing (morphological analysis, morphological disambiguation))。
- 语法的形式化(例如,上下文无关语法,词汇功能语法,中心语驱动的短语结构语法)(grammar formalisms (e.g. CFG, LFG, FUG, HPSG))。
- 针对基于约束的语法编写的词表(lexicons for constraint-based grammars)。
- 计算语义学(semantics)。
- 句子建模与剖析技术(sentence modeling and parsing)。
- 鲁棒的剖析技术(robust parsing)。

4. 语言生成(language generation)

- 句法生成(syntactic generation)。
- 深层生成(deep generation)。

5. 口语输出技术(spoken output technologies)

- 合成语音生成(synthetic speech generation)。
- 用于文本-语音合成(TTS)的文本解释(text interpretation for Text-To-Speech (TTS) Synthesis)。
- 口语生成(从概念到语音)(spoken language generation (conception to speech))。

6. 话语分析与对话(discourse and dialogue)

- 话语建模(discourse modeling)。
- 对话建模(dialogue modeling)。

- 口语对话系统(spoken language dialogue)。

7. 文献自动处理(document processing)

- 文献检索(document retrieval)。
- 文本解释：信息抽取(text interpretation：extracting information)。
- 文本内容的自动归纳(例如，自动文摘)(summarization (e.g. text abstraction))。
- 文本写作和编辑的计算机支持(computer assistance in text creation and editing)。
- 工业和企业中使用的受限语言(controlled languages in industry and company)。

8. 多语问题的计算机处理(multilinguality)

- 机器翻译(machine translation)。
- 人助机译((human-aided) machine translation)。
- 机助人译(machine-aided human translation)。
- 多语言信息检索(multilingual information retrieval)。
- 多语言语音识别(multilingual speech processing)。
- 自动语种验证(automatic language identification)。

9. 多模态的计算机处理 (multimodality)

- 空间和时间的表示方法(从文本中抽取空间和时间的信息)(representations of space and time (automatic abstraction of space and time from text))。
- 文本与图像处理(text and images)。
- 口语与手势的模态结合(使用数据手套)(modality integration：speech and gesture (using data-gloves))。
- 口语与面部信息的模态结合：面部运动与语音识别(modality integration：facial movement & speech recognition)。
- 口语与面部信息的模态结合：面部运动与语音合成(modality integration：facial movement & speech synthesis)。

10. 信息传输与信息存储(transmission and storage)

- 语音编码(语音压缩)(speech coding (speech compression))。
- 语音品质的提升(改善语音的品质)(speech enhancement (speech quality improvement))。

11. 自然语言处理中的数学方法(mathematical methods)

- 统计建模与分类的数学理论(statistical modeling and classification)。
- DSP(数字信号处理)技术(DSP (Digital Signal Processing) techniques)。
- 剖析算法的数学基础研究(parsing techniques)。
- 连接主义的技术(例如，神经网络)(connectionist techniques (e.g. neural network))。
- 有限状态分析技术(finite state technology)。
- 语音和语言处理中的最优化技术和搜索技术(optimization and search in speech and language processing)。

12. 语言资源(language resources)

- 书面语料库(written language corpora)。
- 口语语料库(spoken language corpora)。
- 机器词典与词网的建设(lexicons and word net)。