本文发表于《外国语》（上海外国语大学学报）2011 年第 1 期（总 191 期），p9-17。

计算语言学的历史回顾与现状分析

冯志伟
（教育部语言文字应用研究所，100010）

摘要：本文简要介绍了计算语言学的发展历程，总结了计算语言学中形式模型研究的成果，并分析了当前计算语言学发展的四个特点。

关键词：计算语言学；机器翻译；形式模型；语料库；战略转移

Computational Linguistics: its past and present

Feng Zhiwei
(Institute of Applied Linguistics, Ministry of Education)

Abstract: The author briefly introduces the development process of computational linguistics, summarizes the main success of formal models in computational linguistics, and analyzes four characteristics of current computational linguistics.

Keywords: computational linguistics; machine translation; formal models; corpus; strategic transfer

计算语言学（Computational Linguistics）是当代语言学中的一个新兴学科，在这门学科的发展过程中，曾经在计算机科学、电子工程、语言学、心理学、认知科学等不同的领域分别进行过研究。之所以出现这种情况，是由于计算语言学包括了一系列性质不同而又彼此交叉的学科。本文简要介绍计算语言学的萌芽期、发展期、繁荣期，总结了计算语言学中形式模型研究的成果，并分析了当前计算语言学发展的四个特点。

1. 计算语言学的萌芽期

从 20 世纪 40 年代到 50 年代末这个时期是计算语言学的萌芽期。

在“计算语言学′这个术语出现之前，就有一些具有远见卓识的学者研究过语言的计算问题，他们从计算的角度来研究语言现象，揭示语言的数学面貌。

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》。

1904 年，波兰语言学家 Baudouin de Courtenay 指出，语言学家不仅应当掌握初等数学，而且还要掌握高等数学，他表示坚信，语言学将日益接近精密科学，语言学将根据数学的模式，更多地扩展量的概念，发展新的演绎思想的方法。

1933 年，美国语言学家 L. Bloomfield 提出一个著名的论点：“数学只不过是语言所能达到的最高境界”。

1935 年，加拿大学者 E.Varder Beke 提出了词的分布率的概念，并以之作为词典选词的主要标准。

1944 年，英国数学家 G.U.Yule 发表了《文学词语的统计分析》一书，大规模地使用概率和统计的方法来研究词汇。

这些事实说明，关于语言计算的思想和研究是源远流长的。

有四项基础性的研究特别值得注意：

一项是 Markov 关于马尔可夫模型的研究；

一项是 Turing 关于算法计算模型的研究；

一项是 Shannon 关于概率和信息论模型的研究；

一项是 Chomsky 关于形式语言理论的研究。

早在 1913 年，俄罗斯著名数学家 A. Markov 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用法文发表在俄罗斯皇家科学院的通报上[1]。

后来 A. Markov 的这一思想发展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。

在计算机出现以前，英国数学家 A. M. Turing 就预见到未来的计算机将会对自然语言研究提出新的问题。

1936 年，Turing 向伦敦权威的数学杂志投了一篇论文，题为《论可计算数及其在判定问题中的应用》。在这篇开创性的论文中，Turing 给“可计算性”下了一个严格的数学定义，并提出著名的“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。1950 年 10 月，Turing 在《机器能思维吗》一文中指出：“我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。”

Turing 提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语，进行“Turing 测试”。他天才地预见到计算机和自然语言将会结下不解之缘。

20 世纪 50 年代提出的自动机理论来源于 Turing 在 1936 年提出的可计算性理论和图灵机模型，Turing 的划时代的研究工作被认为是现代计算机科学的基础。Turing 的工作首先导致了 McCulloch-Pitts 的神经元（neuron）理论。一个简单的神经元模型就是一个计算的单元，它可以用命题逻辑来描述。接着，Turing 的工作还导致了 Kleene 关于有限自动机和正则表达式的研究。

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

---

[1] A. A. Markov, Essai d’une recherche statistique sur le texte du roman “Ougene Onegin” illustrant la liaison des epreuve en chain, Bulletin de l’Academie Impériale des Sciences de St-Pétersbourg, 7, 153-162.

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）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[2]。

1956 年，美国语言学家 N. Chomsky 从 Shannon 的工作中吸取了有限状态马尔可夫过程的思想，首先把有限状态自动机作为一种工具来刻画语言的语法，并且把有限状态语言定义为由有限状态语法生成的语言。这些早期的研究工作产生了“形式语言理论”（formal language theory）这样的研究领域，采用代数和集合论把形式语言定义为符号的序列。Chomsky 在研究自然语言的时候首先提出了“上下文无关语法”（Context-free Grammar），后来，Backus 和 Naur 等在描述 ALGOL 程序语言的工作中，分别于 1959 年和 1960 年也独立地发现了这种上下文无关语法。这些研究都把数学、计算机科学与语言学巧妙地结合起来。

Chomsky 在计算机出现的初期把计算机程序设计语言与自然语言置于相同的平面上，用统一的观点进行研究和界说。他在《自然语言形式分析导论》[3]一文中，从数学的角度给语言提出了新的定义，指出：“ 这个定义既适用于自然语言，又适用于逻辑和计算机程序设计理论中的人造语言”。在《语法的形式特性》[4]一文中，他专门用了一节的篇幅来论述程序设计语言，讨论了有关程序设计语言的编译程序问题，这些问题，是作为“组成成分结构的语法的形式研究”，从数学的角度提出来，并从计算机科学理论的角度来探讨的。他在《上下文无关语言的代数理论》[5]一文中提出：“我们这里要考虑的是各种生成句子的装置，它们又以各种各样的方式，同自然语言的语法和各种人造语言的语法二者都有着密切的联系。我们将把语言直接地看成在符号的某一有限集合 V 中的符号串的集合，而 V 就叫做该语言的词汇.....，我们把语法看成是对程序设计语言的详细说明，而把符号串看成是程序。”在这里乔姆斯基把自然语言和程序设计语言放在同一平面上，从数学和计算机科学的角度，用统一的观点来加以考察，对“语言”、“词汇”等语言学中的基本概念，获得了高度抽象化的认识。

Markov, Turing, Shannon 和 Chomsky 这四位著名学者对于语言和计算关系的探讨，是计算语言学萌芽期最重要的研究成果，为计算语言学的理论和技术奠定了坚实的基础。

机器翻译是计算语言学最重要的应用领域。在计算语言学的萌芽期，机器翻译研究得到长足的进展。

1949 年，Weaver 在一篇以《翻译》为题目的《备忘录》[6]中，把机器翻译仅仅看成一种机械的解读密码的过程，他远远没有看到机器翻译在词法分析、句法分析以及语义分析等方面的复杂性。

早期机器翻译系统的研制受到 Weaver 的上述思想的很大影响，许多机器翻译研究者都把机器翻译的过程与解读密码的过程相类比，试图通过查询词典的方法来实现词对词的机器翻译，因而译文的可读性很差，难于付诸实用。

由于学者的热心倡导，实业界的大力支持，美国的机器翻译研究一时兴盛起来。1954 年，美国乔治敦大学在国际商用机器公司（IBM 公司）的协同下，用 IBM-701 计算机，进行了

---

[2] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*，27: pp 379-423, 1948.

[3] N. Chomsky and G. A. Miller, Introduction to the formal analysis of natural languages, In R. D. Luce, R. Bush and E. Galanter, (Eds.) *Handbook of Mathematical Psychology*, Vol. 2. pp. 269-322. Wiley, New York,, 1963.

[4] N. Chomsky, Formal properties of grammars, In R. D. Luce, R. Bush and E. Galanter, (Eds.) *Handbook of Mathematical Psychology*, Vol. 2, pp 323-418, Wiley, New York, 1963.

[5] N. Chomsky and M. P. Schüützenberg, The algebraic theory of context free language [A], In P. Brafford and D. Hirschberger, Computer Programming and Formal Language [C], Amsterdam, North Holland, pp. 118-161.

[6] 参看 W. N. Locke and A. D. Booth (eds.), Machine translation of languages: fourteen essays, Cambridge, Mass: Technology Press of the Massachusetts Institute of Technology, 1955.

世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

1952 年，在美国的 MIT 召开了第一次机器翻译会议，在 1954 年，出版了第一本机器翻译的杂志，这个杂志的名称就叫做 Machine Translation（《机器翻译》）。尽管人们自然语言的计算方面进行了很多的研究工作，但是，直到 20 世纪 60 年代中期，才出现了 computational linguistics（计算语言学）这个术语，而且，在刚开始的时候，这是术语是偷偷摸摸地、羞羞涩涩地出现的。

1965 年 Machine Translation 杂志改名为 Machine Translation and Computational Linguistics（《机器翻译和计算语言学》）杂志，在杂志的封面上，首次出现了“Computational Linguistics”这样的字眼，但是，“and Computational Linguistics”这三个单词是用特别小号的字母排印的。这说明，人们对于“计算语言学”是否能够算为一门真正的独立的学科还没有把握。计算语言学刚刚登上学术这个庄严的殿堂的时候，还带有“千呼万唤始出来，犹抱琵琶半遮面”那样的羞涩，以至于人们不敢用 Machine Translation 同样大小的字母来排印它。当时 Machine Translation 杂志之所以改名，是因为在 1962 年美国成立了“机器翻译和计算语言学学会”（Association for machine Translation and Computational Linguistics），通过改名可以使杂志的名称与学会的名称保持一致。

根据这些史料，我们认为，远在 1962 年，就出现了“计算语言学”这个学科了，尽管它在刚出现的时候还是偷偷摸摸的，显示出少女般的羞涩。但是，无论如何，计算语言学这个新兴的学科终于萌芽了，她破土而出，悄悄地登上了学术的殿堂。

1964 年，美国科学院成立了语言自动处理谘询委员会（Automatic Language Processing Advisory Committee，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[7]，这个报告对机器翻译采取了否定的态度，报告宣称：“在目前给机器翻译以大力支持还没有多少理由”；这个报告还指出，机器翻译研究遇到了难以克服的“语义障碍”（semantic barrier）。在 ALPAC 报告的影响下，许多国家的机器翻译研究低潮，许多已经建立起来的机器翻译研究单位遇到了行政上和经费上的困难，在世界范围内，机器翻译的热潮突然消失了，出现了空前萧条的局面。

美国语言学家 David Hays 是 ALPAC 委员会的成员之一，他参与起草了 ALPAC 报告，在 ALPAC 报告中，他建议，在放弃机器翻译这个短期的工程项目的时候，应当加强语言和自然语言计算机处理的基础研究，可以把原来用于机器翻译研制的经费使用到自然语言处理的基础研究方面，David Hays 把这样的基础研究正式命名为 Computational Linguistics（计算语言学）。所以，我们可以说，“计算语言学”这个学科名称最早出现于 1962 年，而在 1966 年才在美国科学院的 ALPAC 报告中正式得到学术界的承认。

2. 计算语言学的发展期

20 世纪 60 年代中期到 80 年代末期年是计算语言学的发展期。

在计算语言学的发展期，各个相关学科的彼此协作，联合攻关，取得了一些令人振奋的成绩。

统计方法在语音识别算法的研制中取得成功。其中特别重要的是“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）。这些模型是分别独立地由两支队伍研制的。一支是 Jelinek，Bahl，Mercer 和 IBM 的华生研究中心的研究人员，另一支是卡内基梅隆大学（Carnegie Mellon University）的 Baker

[7] ALPAC, Language and machines: computer in translation and linguistics, A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Publication 1416, Washington.

等，Baker 受到普林斯顿防护分析研究所的 Baum 和他的同事们的工作的影响。AT&T 的贝尔实验室（Bell laboratories）也是语音识别和语音合成的中心之一。

逻辑方法在计算语言学中取得了很好的成绩。1970 年，Colmerauer 和他的同事们使用逻辑方法研制了 Q 系统（Q-system）和“变形语法”（metamorphosis grammar）并在机器翻译中得到应用，Colmerauer 还是 Prolog 语言的先驱者，他使用逻辑程序设计的思想设计了 Prolog 语言。1980 年 Pereira 和 Warren 提出的“定子句语法”（Definite Clause Grammar）也是在计算语言学中使用逻辑方法的成功范例之一。1979 年 Kay 对于“功能语法”（functional grammar）的研究，1982 年 Bresnan 和 Kaplan 在“词汇功能语法”（Lexical Function Grammar，简称 LFG）方面的工作，都是特征结构合一（feature structure unification）研究方面的重要成果，他们的研究引入了“复杂特征”（complex feature）的概念，与此同时，我国学者冯志伟提出了“多叉多标记树形图模型”（Multiple-branched Multiple-labeled Tree Model，简称 MMT 模型）[8]，在他设计的多语言机器翻译 FAJRA（ 英语、法语、日语、俄语、德语的法文首字母缩写）系统中，采用了“多标记”（Multiple label）的概念。“多标记”的概念与“复杂特征”的概念实质上是一致的，这些关于自然语言特征结构研究成果，都有效地克服了 Chomsky 短语结构语法的生成能力过强的缺陷。

在这个时期，自然语言理解（natural language understanding）也取得明显的成绩。自然语言理解肇始于 Terry Winograd 在 1972 年研制的 SHRDLU 系统，这个系统能够模拟一个嵌入玩具积木世界的机器人的行为。该系统的程序能够接受自然语言的书面指令（例如，“Move the red block on top of the smaller green one”[请把绿色的小积木块移动到红色积木块的上端]），从而指挥机器人摆弄玩具积木块。这是一个非常复杂而精妙的系统。这个系统还首次尝试建立基于 Halliday 系统语法（systemic grammar）的全面的英语语法。Winograd 的模型还清楚地说明，句法剖析也应该重视语义和话语的模型。1977 年，Roger Schank 和他在耶鲁大学的同事和学生们建立了一些语言理解程序，这些程序构成一个系列，他们重点研究诸如脚本、计划和目的这样的人类的概念知识以及人类的记忆机制。他们的工作经常使用基于网络的语义学理论，并且在他们的表达方式中开始引进 Fillmore 在 1968 年提出的关于“深层格”（deep case）的概念。

在自然语言理解研究中也使用过逻辑学的方法，例如 1967 年 Woods 在他研制的 LUNAR 问答系统中，就使用谓词逻辑来进行语义解释。

计算语言学在话语分析（discourse analysis）方面也取得了很大的成绩。基于计算的话语分析集中探讨了话语研究中的四个关键领域：话语子结构的研究、话语焦点的研究、自动参照消解的研究、基于逻辑的言语行为的研究。1977 年，Crosz 和她的同事们研究了话语中的“子结构”（substructure）和话语焦点；1972 年，Hobbs 开始研究“自动参照消解”（automatic reference resolution）。在基于逻辑的言语行为研究中，Perrault 和 Allen 在 1980 年建立了“信念－愿望－意图”（Belief-Desire-Intention，简称 BDI）的框架。

在 1983－1993 年的十年中，计算语言学研究者对于过去的研究历史进行了反思，发现过去被否定的有限状态模型和经验主义方法仍然有其合理的内核。在这十年中，计算语言学的研究又回到了 20 世纪 50 年代末期到 60 年代初期几乎被否定的有限状态模型和经验主义方法上去，之所以出现这样的复苏，其部分原因在于 1959 年 Chomsky 对于 Skinner 的“言语行为”（Verbal Behavior）的很有影响的评论在 80 年代和 90 年代之交遭到了理论上的反对。

这种反思的第一个倾向是重新评价有限状态模型，由于 Kaplan 和 Kay 在有限状态音系学和形态学方面的工作，以及 Church 在句法的有限状态模型方面的工作，显示了有限状态模型仍然有着强大的功能，因此，这种模型又重新得到计算语言学界的注意。

这种反思的第二个倾向是所谓的“重新回到经验主义”；这里值得特别注意的是语音和

[8] 冯志伟，汉语句子的多叉多标记树形图分析法，《人工智能学报》，1983 年，第 2 期。

语言处理的概率模型的提出，这样的模型受到 IBM 公司华生研究中心的语音识别概率模型的强烈影响。这些概率模型和其他数据驱动的方法还传播到了词类标注、句法剖析、介词短语附着歧义的判定以及从语音识别到语义学的联接主义方法的研究中去。

此外，在这个时期，自然语言的生成研究也取得了引人瞩目的成绩。

3. 计算语言学的繁荣期

从 20 世纪 90 年代开始，计算语言学进入了繁荣期。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，自 1989 年以来，机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是，在基于规则的技术中引入了语料库方法，其中包括统计方法，基于实例的方法，通过语料加工手段使语料库转化为语言知识库的方法，等等。这种建立在大规模真实文本处理基础上的机器翻译，是机器翻译研究史上的一场革命，它将会把计算语言学推向一个崭新的阶段。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。

在 20 世纪 90 年代的最后五年（1994-1999），计算语言学的研究发生了很大的变化，出现了空前繁荣的局面。这主要表现在如下三个方面。

第一，概率和数据驱动的方法几乎成为了计算语言学的标准方法。句法剖析、词类标注、参照消解、话语处理、机器翻译的算法全都开始引入概率，并且采用从语音识别和信息检索中借过来的基于概率和数据驱动的评测方法。

第二，由于计算机的速度和存储量的增加，使得在计算语言学的一些应用领域，特别是在语音合成、语音识别、文字识别、拼写检查、语法检查这些应用领域，有可能进行商品化的开发。自然语言处理的算法开始被应用于“增强交替通信”（Augmentative and Alternative Communication，简称 AAC）中，语音合成、语音识别和文字识别的技术被应用于“移动通信”（mobile communication）中。

第三，随着网络技术的发展，互联网（Wide World Web）逐渐变成一个多语言的网络世界，互联网上的机器翻译、信息检索和信息抽取的需要变得更加紧迫。目前，在互联网上除了使用英语之外，越来越多地使用汉语、西班牙语、葡萄牙语、德语、法语、俄语、日语、韩国语等英语之外的语言。从 2000 年到 2005 年，互联网上使用英语的人数仅仅增加了 126.9%，而在此期间，互联网上使用俄语的人数增加了 664.5%，使用葡萄牙语的人数增加了 327.3%，使用中文的人数增加了 309.6%，使用法语的人数增加了 235.9%。

2008 年 6 月，中国的网民已经达到 2.56 亿，超过了美国的网民数量，截至 2010 年 5 月，我国网民的数量已经达到 4.04 亿之多，使用手机上网的网民达到 2.33 亿人，我国成为了世界上首屈一指的互联网大国。截至 2009 年，我国共完成互联网基础设施建设投资 4.3 万亿元，建成光缆网络线路总长度达 826.7 万公里。目前，我国 99.1％的乡镇和 92％的行政村接通了互联网，95.6％的乡镇接通了宽带，3G 网络已基本覆盖全国。2009 年我国电子商务交易总额突破 4 万亿元。互联网已经成为我国经济发展的火车头。

互联网上使用英语之外的其他语言的人数增加得越来越多，英语在互联网上独霸天下的局面已经打破，互联网确实已经变成了多语言的网络世界，因此，网络上的不同自然语言之间的计算机自动处理也就变得越来越迫切了。网络上的机器翻译、信息获取和信息搜索正在迅猛地发展，计算语言学的各种应用技术事实上已经成为了互联网技术的重要支柱。

在信息时代，科学技术的发展日新月异，新的信息、新的知识如雨后春笋地不断增加，出现了“信息爆炸”（information explosion）的局面。现在，世界上出版的科技刊物达 165000 种，平均每天有大约 2 万篇科技论文发表。专家估计，我们目前每天在互联网上传输的数据量之大，已经超过了整个 19 世纪的全部数据的总和；我们在新的 21 世纪所要处理的知识总

量将要大大地超过我们在过去 2500 年历史长河中所积累起来的全部知识总量。而所有的这些信息主要都是以语言文字作为载体的，也就是说，网络世界主要是由语言文字构成的。

为了说明计算语言学的重要性，我们可以把它与物理学做如下的类比：**我们说物理学之所以重要，是因为物质世界是由物质构成的，而物理学恰恰是研究物质运动的学科；我们说计算语言学之所以重要，是因为网络世界主要是由语言文字构成的，而计算语言学恰恰是研究语言文字自动处理的学科。**

可以预见，知识突飞猛进的增长和网络技术日新月异的进步，一定会把计算语言学的研究推向一个崭新的阶段。计算语言学有可能成为当代语言学中最有发展潜力的学科，计算语言学已经给有着悠久传统的古老的语言学注入了新的生命力，在计算语言学的推动下，语言学有可能真正成为当代科学百花园中的一门名副其实的领先学科。

4. 计算语言学中形式模型的研究

计算语言学有着明确的应用目标，语音合成、语音识别、信息检索、信息抽取、机器翻译等，都是计算语言学的重要应用领域。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，在这众多的应用领域中，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），建立自然语言的“计算模型”（computational model），使之能够在计算机上实现。在计算语言学中，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。因此，这种建立语言形式模型的研究是非常重要的，它应当属于计算语言学的基础理论研究。

由于自然语言的复杂性，这样的形式模型的研究往往是一个“**强不适定问题**”（strongly ill-posed problem），也就是说，在用形式模型建立算法来求解计算语言学的问题时，往往难以满足问题解的“**存在性**”、“**唯一性**”和“**稳定性**”的要求，有时是不能满足其中的一条，有时甚至三条都不能满足。因此，对于这样的强不适定性问题求解，应当加入适当的“**约束条件**”（constraint conditions），使问题的一部分在一定的范围内变成“适定问题”（well-posed problem），从而顺利地求解这个问题。

计算语言学是一个多边缘的交叉学科，因此，我们可以通过计算机科学、语言学、心理学、认知科学、人工智能等多学科的通力合作，把人类知识的威力与计算机的计算能力结合起来，给计算语言学的形式模型提供大量的、丰富的“约束条件”，从而解决计算语言学的各种困难问题。计算语言学这个学科的边缘性、交叉性的特点，为解决这样的“强不适定问题”提供了有力的手段，我们有可能把计算语言学形式模型的研究这个“强不适定问题”变成“适定问题”，这是我们在研究计算语言学的形式模型的时候，值得特别庆幸的，也是应该特别注意的。

早在计算语言学这个学科出现之前，语言计算研究的先驱者们就开始探索自然语言的形式模型。例如，Markov 链，Zipf 定律，Shannon 关于“熵”的研究，Bar-Hillel 的范畴语法，Harris 的语言串分析法，О．С．Кулагина 的语言集合论模型等。Markov 等具有远见卓识的学者很早就从形式描述的角度来研究自然语言，开**计算语言学形式模型**（Formal models for NLP）研究的先河。

随着计算语言学研究的发展，一系列的形式模型开始建立起来。这些形式模型大致可以归纳为如下几种[9]：

[9] 冯志伟，自然语言处理的形式模型，中国科学技术大学校友文库，中国科学技术大学出版社，2009 年。

- 基于短语结构语法的形式模型：主要有 Chomsky 的短语结构语法，递归转移网络和扩充转移网络，自底向上分析法与自顶向下分析法，通用句法生成器和线图分析法，Earley 算法，左角分析法，CKY 算法，Tomita 算法，Chomsky 的管辖-约束理论与最简方案，Joshi 的树邻接语法等。
- 基于合一运算的形式模型：主要有 Kaplan 的词汇功能语法，Kay 的功能合一语法，Gazdar 的广义短语结构语法，Shieber 的 PATR，Pollard 的中心语驱动的短语结构语法，Pereira 的定子句语法等。
- 基于依存和配价的形式模型：主要有 Tesnière 的依存语法，德国学者的配价语法，Hudson 的词语法等。
- 基于格语法的形式模型：主要有 Fillmore 的格语法和框架网络。
- 基于词汇主义的形式模型：主要有 Gross 的词汇语法，Sleator 和 Temperley 的链语法，Baldridge 等的组合式范畴语法（Combinatory Category Grammar, 简称 CCG），词网（WordNet）等。
- 基于概率和统计的形式模型：主要有 N-元语法（N-gram），隐马尔可夫模型（Hidden Markov Model，简称 HMM），最大熵模型（Maximum Entropy，简称 ME），条件随机场（Condition Random Field, 简称 CRF），Charniak 的概率上下文无关语法和词汇化的概率上下文无关语法，Bayes 公式，动态规划算法，噪声信道模型，最小编辑距离算法，决策树模型，加权自动机，Viterbi 算法，向内向外算法，向前向后算法等。
- 语义自动处理的形式模型：主要有义素分析法、语义场理论，语义网络理论，Montague 的蒙塔鸠语法，Wilks 的优选语义学，Schank 的概念依存理论，Mel'chuk 的意义-文本理论等。
- 语用自动处理的形式模型：主要有 Mann 和 Thompson 的修辞结构理论，文本连贯中的常识推理技术等。

计算语言学形式模型的研究大大地丰富了传统的理论语言学的内容，是计算机时代理论语言学研究的重要成果，我们应当特别关注这个领域的研究。

5. 当前计算语言学发展的特点

21 世纪以来，由于互联网的普及，自然语言的计算机处理成为了从互联网上获取知识的重要手段，生活在信息网络时代的现代人，几乎都要与互联网打交道，都要或多或少地使用计算语言学的研究成果来帮助他们获取或挖掘在广阔无边的互联网上的各种知识和信息，因此，世界各国都非常重视计算语言学的研究，投入了大量的人力、物力和财力。

当前国外计算语言学研究有四个显著的特点：

第一，**随着语料库建设和语料库语言学的崛起，大规模真实文本的处理成为计算语言学的主要战略目标，计算语言学中出现了“战略转移”**（strategic transit）：在过去的五十多年中，从事计算语言学系统开发的绝大多数学者，都把自己的目的局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法-语义分析，尽管这些应用系统在某些受限的“子语言”（sub-language）中也曾经获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难。因为从自然语言系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往的任何系统所远远不及的。而且，随着系统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表示和管理知识等基本问题上，不得不另辟蹊径。这样，就提出了大规模真实文本的自动处理问题。1990年8月在芬兰赫尔辛基举行的第13届国际计算语言学会议（即COLING'90）为会前讲座确定的主题是：“处理大规模真实文本的理论、方法和工具”，这说明，实现大规

模真实文本的处理将是计算语言学在今后一个相当长的时期内的战略目标。为了实现战略目标的转移，需要在理论、方法和工具等方面实行重大的革新。1992年6月在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议（TMI-92）上，宣布会议的主题是“机器翻译中的经验主义和理性主义的方法”。所谓“**理性主义**”(rationalism)，就是指以生成语言学为基础的方法，所谓“**经验主义**”（empiricism），就是指以大规模语料库的分析为基础的方法。从中可以看出当前计算语言学关注的焦点。当前语料库的建设和语料库语言学的崛起，正是计算语言学战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的分析方法（即经验主义的方法）至少是对基于规则的分析方法（即理性主义的方法）的一个重要补充。因为从“大规模”和“真实”这两个因素来考察，语料库才是最理想的语言知识资源。但是，要想使语料库名符其实地成为自然语言的知识库，就有必要首先对语料库中的语料进行自动标注，使之由“生语料”变成“熟语料”，以便于人们从中提取丰富的语言知识。

第二，**计算语言学中越来越多地使用机器自动学习的方法来获取语言知识**：传统语言学基本上是通过语言学家归纳总结语言现象的手工方法来获取语言知识的，由于人的记忆能力有限，任何语言学家，哪怕是语言学界的权威泰斗，都不可能记忆和处理浩如烟海的全部的语言数据，因此，使用传统的手工方法来获取语言知识，犹如以管窥豹，以蠡测海，这种获取语言知识的方法带有很大的主观性。传统语言学中啧啧地称道的所谓“例不十，不立法；例外不十，法不破”[10]的朴学精神，貌似严格，实际上，在浩如烟海的语言数据中，以十个例子或十个例外就轻而易举地来决定语言规则的取舍，难道就能够万无一失地保证这些规则是可靠的吗？这是大大地值得怀疑的。当前的计算语言学研究提倡建立语料库，使用机器学习的方法，让计算机自动地从浩如烟海的语料库中获取准确的语言知识。机器词典和大规模语料库的建设，成为了当前计算语言学的热点。这是语言学获取语言知识方式的巨大变化，作为 21 世纪的语言学工作者，应该注意到这样的变化，逐渐改变传统的获取语言知识的手段。

第三，**计算语言学中越来越多地使用统计数学方法来分析语言数据**：使用人工观察和内省的方法，显然不可能从浩如烟海的语料库中获取精确可靠的语言知识，必须使用统计数学的方法。目前，计算语言学中的统计数学方法已经相当成熟，如果我们认真地学会了统计数学，努力地掌握了统计数学，就会使我们在获取语言知识的过程中如虎添翼。目前，在机器翻译中使用统计方法获得了很好的成绩，统计机器翻译[11]（statistical machine translation，简称 SMT）成为了机器翻译的主流技术。

2003 年 7 月，在美国马里兰州巴尔的摩（Baltimore, Maryland）由美国商业部国家标准与技术研究所 NIST/TIDES (National Institute of Standards and Technology) 主持的评比中，来自德国亚琛大学（Achen University）的年青的博士研究生 F. J. Och 获最好成绩。他使用统计方法，在很短的时间之内就构造了阿拉伯语和汉语到英语的若干个机器翻译系统。两千多年前，伟大的希腊科学家 Archimedes（阿基米德）说过：“只要给我一个支点，我就可以移动地球。”(“Give me a place to stand on, and I will move the world.”)，而这次评比中，Och 也模仿着 Archimedes 说：“只要给我充分的并行语言数据，那么，对于任何的两种语言，我就可以在几小时之内给你构造出一个机器翻译系统。”(“Give me enough parallel data, and you can have translation system for any two languages in a matter of hours.”) 这反映了新一代的机器翻译研究者朝气蓬勃的探索精神和继往开来的豪情壮志。看来，Och 似乎已经找到了机器翻译的有效方法，至少按照他的路子走下去，也许有可

[10]王力在《汉语史稿》（上册）（1980）中指出，“所谓区别一般与特殊，那是辩证法的原理之一。在这里我们指的是黎锦熙先生所谓‘例不十，不立法’。我们还要补充一句，就是‘例外不十，法不破’。”

[11] Philipp Koehn, Statistical machine translation, Cambridge University Press, 2010.

能开创出机器翻译研究的一片新天地，使我们在探索真理的曲折道路上看到了耀眼的曙光。过去我们研制一个机器翻译系统往往需要几年的时间，而现在采用 0ch 的方法构造机器翻译系统只要几个小时就可以了，研制机器翻译系统的速度已经大大地提高了。这是当前计算语言学中令人兴奋的新进展。

第四，**计算语言学中越来越重视词汇的作用，出现了“词汇主义”（lexicalism）的倾向**：词汇信息在自然语言的计算机处理中起着举足轻重的作用，单词之间的相似度（similarity）的计算、词汇搭配关系（lexical collocation）和词汇联想关系（lexical association）的自动获取、动词的次范畴框架（sub-categorization frame of verb）的自动获取、计算词汇语义学（computational lexical semantics）等都是当前计算语言学研究的热点。在统计方法中引入了词汇信息，可以大大地提高统计分析的精确度，在句法分析中引入词汇信息，可以减少结构上歧义，提高句法分析的效率。机器可读词典和词汇知识库成为了自然语言处理最关键、最重要的语言资源。

6. 大哉计算语言学之为用

现在计算语言学正处于激动人心的时刻。普通计算机用户可以使用的计算资源正以惊人的速度迅速增长，互联网的兴起并且成为了无比丰富的信息资源，无线移动通信日益普及并且日益增长起来，这些都使得计算语言学的应用成为了当前科学技术的热门话题。

这里我想列举出计算语言学一些当前的应用项目，由此可以看出这个学科近期发展对于社会进步的重要作用。

- **自动生成天气预报**：加拿大的计算机程序 TAUM-METEO 能够接受每天的天气预报的数据，然后自动生成天气预报的报告，不必经过进一步的编辑就可以用英语和法语公布。
- **自动翻译和自动问答**：美国 Systran 的 Babel Fish 机器翻译系统每天可以从 Alta Vista 搜索引擎处理 100 万个翻译的问题。基于网络的问答系统（Web-based question answering）是简单的网络搜索的进一步发展，在基于网络的问答系统中，用户不只是仅仅键入关键词进行提问，而是可以用自然语言提出一系列完整的问题，从容易的问题到困难的问题都可以提，计算机根据网络搜索的结果，用自然语言回答用户的提问。
- **饭馆咨询服务**：目前，世界上已经出现不少使用自然语言的口语向计算机咨询饭馆服务情况的系统。例如，前往美国 Massachusetts 州 Cambridge 访问的一个访问者用口语问计算机在什么地方可以吃饭。系统查询了一个关于当地饭馆的数据库之后，给出有关信息用自然语言做出回答。
- **图象到语音的自动转换**：给计算机装上图象识别系统，它就可以观看一段足球比赛的录像，并且用自然语言实时地向足球爱好者报告比赛的情况。
- **残疾人增强交际**：对于有言语或交际障碍的残疾人，计算机能预见到在说话过程中下面将要出现的词语，给他们做出提示，或者帮助他们说话时在词语方面进行扩充，使残疾人能完整地说出简洁的话语。
- **旅行咨询服务**：例如，美国的 Amtrak 旅行社、美国联合航空公司以及其他的一些旅行社可以与智能会话代理（intelligent conversation agent）进行交互，在智能会话代理的指导下，他们能够自动地处理关于旅行中的订票、到达、离开等方面的信息。
- **语音地理导航**：汽车制造公司可以给汽车驾驶员提供语音识别和文本-语音转换系统，使得他们可以通过语音来控制他们的环境、娱乐以及导航系统，从而可以自由地使用他们的双手操纵汽车。在国际空间站的宇航员也可以使用简单的口语对话系统来帮助他们的工作。语音合成系统还可以作为全球定位系统（Global Positioning System，简称 GPS）的语音导航，使用自动合成的语音来报告地理情况，保证驾驶员用双手操纵

汽车。目前使用语音导航的 GPS 已经逐渐普及，给汽车驾驶员提供了极大的方便。

- **语音资料搜索**：一些视频搜索公司使用语音识别技术，可以在网络上提供多达数百万小时的视频资料的搜索服务，并且在语音资料中搜索到与之相应的单词。
- **跨语言信息检索和翻译**：Google（谷歌）在网上提供跨语言信息检索和 40 多个语言对的自动翻译服务，用户可以使用他们自己的母语来提问，以便搜索其他语言中的有关信息。Google 还可以对用户提出的问题进行自动翻译，找出与所提出的问题最相关的网页，然后自动地把它们翻译成用户的母语。
- **作文自动评分**：在美国，像 Pearson（培生公司）这样的大型出版社和像 ETS（English Test Service）这样的测试服务公司使用自动系统来分析数千篇学生的英语作文，对于这些作文进行自动打分、自动排序和自动评价，而且计算机的打分结果与人的打分结果几乎毫无二致，难以分辨。
- **自动阅读家庭教师**：让计算机充当自动阅读家庭教师，帮助改善阅读能力，它能教小孩阅读故事。当阅读人要求阅读或者出现阅读错误时，计算机能使用语音识别器来进行干预。具有生动活泼的动画特征的交互式虚拟智能代理可以充当教员来教儿童学习如何阅读。
- **个性化市场服务**：文本分析公司根据用户在互联网论坛和用户群体组织中表现出来的意见、偏好、态度的自动测试结果，对用户提供智能化、个性化的服务，帮助用户在市场上挑选到符合他们要求的商品。

计算语言学这些应用项目的成就确实是鼓舞人心的。我们情不自禁地赞叹：**“大哉计算语言学之为用！”**

我国计算语言学已经取得不少成绩，但是，与国际水平相比，差距还很大。2010 年 8 月 23 日-27 日第 23 届国际计算语言学会议在北京召开，与会代表 700 多人，这说明我国的计算语言学研究已经引起了国际计算语言学界的广泛关注。计算语言学是国际性的学科，我们不仅要学习和了解国外计算语言学的研究成果和最新动态，而且要参与到国际计算语言学的研究中去，用国际的水平和国际的学术规范来要求我们的研究，促进我国计算语言学研究的世界化。

[作者简介] 冯志伟，1939 年生，教育部语言文字应用研究所研究员、博士生导师，研究方向为计算语言学、语料库语言学。

参考文献

[1] 冯志伟，自然语言的计算机处理[M]，上海，上海外语教育出版社，1996 年。
[2] 冯志伟，机器翻译研究[M]，北京，中国对外翻译出版公司，2004 年 12 月。
[3] 冯志伟，自然语言处理的形式模型[M]，中国科学技术大学出版社，2010 年。
[4] 冯志伟，语言与数学，世界图书出版公司，2010 年。
[5] Bill Manaris, Natural language processing: A human-computer interaction perspective [A], Advances in Computers, Volume 47，1999.
[6] Carstensen Kai-Uwe et al, Computerlinguistik und Sprachtechnologie, Eine Einführung [M], Heidelberg/Berlin, Spektrum Akademischer Verlag, 2004.
[7] Daniel Jurafsky, James H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition [M], Upper Saddle River, New Jersey, Prentice Hall, 2000.。中文译本，冯志伟、孙乐 译，《自然语言处理综论》，电子工业出版社，2005 年。
[8] Philipp Koehn, Statistical Machine Translation [M], Cambridge University Press, 2010.