自然语言处理的学科定位

冯志伟

(教育部语言文字应用研究所，北京 100010)

摘 要：自然语言处理是当代语言学中的一个重要学科，对它进行正确的学科定位，确定它在学科体系中的位置，有助于推动它的发展。本文从自然语言处理的过程、范围以及历史三个角度来考察它的学科定位问题，论证了自然语言处理是一个多边缘的交叉学科，其研究以语言学为主，涉及计算机科学、数学、心理学、哲学、逻辑学、统计学、电子工程、生物学各个领域。我们应当努力更新知识，以适应自然语言处理的要求。

关键词：自然语言处理；学科定位；交叉学科

中图分类号：H087 **文献标识码**：A **文章编号**：1002-722X（2005）03-0001-08

Academic Position of Natural Language Processing

FENG Zhi-wei

(Institute of Applied Linguistics, Ministry of Education, Beijing, 100010, China)

Abstract: Natural Language Processing (NLP) is an important section in contemporary linguistics. Defining the position of NLP in the academic system will be very helpful for the development of NLP. This paper tries to define the academic position of NLP from three aspects: its procedure, scope and development. It is concluded that NLP is a multilateral cross discipline, with linguistics as the main part, and also in relation to computer science, mathematics, psychology, philosophy, logic, statistics, electronic engineering, and biology. We have to make great efforts to update our knowledge, in order to meet the challenge of NLP.

Key words: Natural Language Processing; academic position; cross discipline

0. 引言

采用计算机技术来研究和处理自然语言是 20世纪 40年代末期和 50年代才开始的，50多年来，这项研究取得了长足的进展，成为了当代语言学中一门重要的新兴学科——自然语言处理（Natural Language Processing，简称 NLP）。在信息网络时代，自然语言处理引起了越来越多的语言学者的重视，成为了当代语言学中的“显学”。如何对自然语言处理进行正确的学科定位，使我们认识到它在学科体系中的位置，从而自觉地推动其发展，是一个至关重要的问题。

我们可以从自然语言处理的过程、范围以及历史三个角度——即从共时和历时两个层面——来考察它的学科定位问题。

1. 自然语言处理的过程

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：

第一，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；第二，把这种严密而规整的数学形式表示为算法，使之在计算上形式化；第三，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；第四，对于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。

美国计算机科学家 Bill Manaris在 1999年出版的《计算机进展》（*Advanced in Computers*）第 47卷的《从人机交互的角度看自然语言处理》一文中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和

收稿日期：2005 - 02 - 25

作者简介：冯志伟（1939 - ），男，云南昆明人，教育部语言文字应用研究所研究员，博士生导师，语言学和信息科学双硕士学位，研究方向为计算语言学、自然语言处理、机器翻译及应用语言学。

语言应用的模型，建立计算框架来实现这样的语言模型，提出相应的方法不断地加以完善，根据模型设计各种实用系统，并探讨这些实用系统的评测技术。

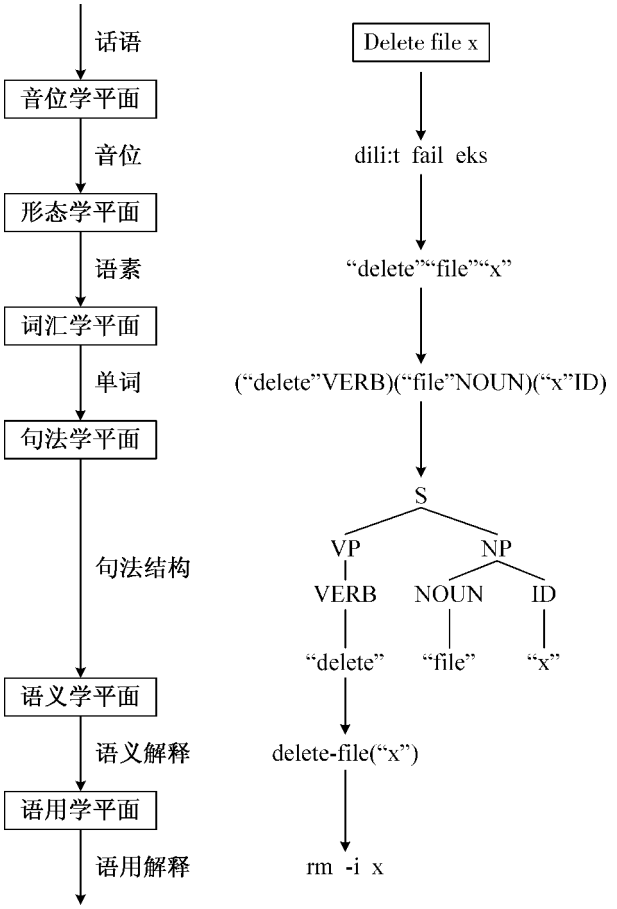
Bill Manaris关于自然语言处理的这个定义，比较全面地表达了计算机对自然语言的研究和处理的上述四个方面的过程。我们认同这样的定义。

根据这样的定义，我们认为，建立自然语言处理模型需要如下不同平面的知识：

（一）声学和韵律学的知识；（二）音位学的知识；（三）形态学的知识；（四）词汇学的知识；（五）句法学的知识；（六）语义学的知识；（七）话语分析的知识；（八）语用学的知识；（九）外部世界的常识性知识。

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，大多数研究人员都认为，这些知识至少可以分为词汇学、句法学、语义学和语用学等平面。每一个平面传达信息的方式各不相同。下面具体说明在自然语言处理中这些知识平面的一般情况。如果要让自然语言处理系统理解并执行口头指令“Delete file x”（“删除文件X”），一般来说需要经过如下的处理过程：

图 1 自然语言处理系统中的知识平面



从图 1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“/dili:t 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。

这个例子来自美国学者 Wilensky为 UNIX设计的一个语音理解界面，叫做 UNIX Consultant。这个界面使用了上述第 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个方面的知识；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。但是，这些知识是要通过计算机来实现和完成的，需要建立数学模型，进行算法设计和逻辑推理，还需要心理学、哲学、逻辑学和生物学提供理论和方法，如果要预测统计事件，还需要统计学的知识，如果要做语音输入和输出，还需要使用信号处理的技术，因此，除了语言学之外，自然语言处理系统还要涉及如下的知识领域：

（一）计算机科学：提供模型表征、算法设计和计算机实现的技术；（二）数学：提供形式化的数学模型和形式化的数学方法；（三）心理学：提供人类言语行为的心理模型和理论；（四）哲学：提供关于人类的思维和语言的更深层次的理论；（五）逻辑学：提供逻辑运算和逻辑推理的理论和方法；（六）统计学：提供基于样本数据来预测统计事件的技术；（七）电子工程：提供信息论的理论基础和语言信号处理技术；（八）生物学：提供大脑中人类语言行为机制的理论。

由此可见，自然语言处理是一个多边缘的交叉学科，它的研究必须结合各边缘学科的知识。每个从事自然语言处理研究的人，都应该进行更新知识的再学习，尽量使自己成为文理兼通、博学多识的人。当然，一个人很难精通上述各个领域的知识，但是，至少在他自己的专业领域应该是博贯精通的内行，对于相关的领域不是似懂非懂的外行，这样才有可能得心应手地进行自然语言处理的研究工作。

2. 自然语言处理的范围

下面我们从自然语言处理的范围，也就是从横的角度来考察它的学科定位。

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：

(一) 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。这个方向最重要的研究领域是语法形式化理论和数学理论；(二) 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。这一方向早期的研究有术语数据库的建设、各种机器可读的电子词典的开发，近年来则有大规模语料库的涌现；(三) 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；(四) 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, 简称 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 und 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项内容：

(一) 口语输入：语音识别、信号表征(语音信号分析)、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、语言表征理论(语言模型)、说话人识别、口语理解；

(二) 书面语输入：文献格式识别、光学字符识别(印刷体及手写体)、手写界面(如用笔输入的计算机)、手写文字分析(如签名验证)；

(三) 语言分析和理解：小于句子单位的处理(形态分析，形态排歧)、语法的形式化(如上下文无关语法、词汇功能语法、功能合一语法、中心语驱动的短语结构语法)、基于约束语法的词表(Lexicons for Constraint-Based Grammars)、计算语义学、句子建模与剖析技术、鲁棒的剖析技术(Robust Parsing)；

(四) 语言生成：句法生成、深层生成；

(五) 口语输出技术：合成语音生成、用于文本—语音合成的文本解释(Text Interpretation for Text-to-Speech Synthesis)、口语生成(从概念到语音)(Spoken Language Generation: Conception to Speech)；

(六) 话语分析与对话：话语建模(Discourse Modeling)、对话建模、口语对话系统；

(七) 文献自动处理：文献检索、文本解释：信息抽取、本文内容的自动归纳(如自动文摘)、文本写作和编辑的计算机支持、工业和企业中使用的受限语言(Controlled Languages in Industry and Company)；

(八) 多语问题的计算机处理：机器翻译、人助机译、机助人译、多语言信息检索、多语言语音识别、自动语种验证；

(九) 多模态的计算机处理：空间和时间的表征方法(从文本中抽取空间和时间的信息)、文本与图像处理、口语与手势的模态结合(使用数据手套)、口语与面部信息的模态结合：面部运动与语音识别、口语与面部信息的模态结合：面部运动与语音合成；

(十) 信息传输与信息存储：语音编码(语音压缩)、语音品质提升；

(十一) 自然语言处理中的数学方法：统计建模与分类的数学理论、DSP(数字信号处理)技术、剖析算法的数学基础研究、连接主义的技术(如神经网络)、有限状态分析技术、语音和语言处理中的最优化技术和搜索技术；

(十二) 语言资源：书面语料库、口语语料库、机器词典与词网的建设、术语编纂与术语数据库、网络数据挖掘与信息提取；

(十三) 自然语言处理系统的评测：面向任务的文本分析评测、机器翻译系统和翻译工具的评

测、大覆盖面的自然语言剖析器的评测、人的因素与用户的可接受性、语音识别：评估与评测、语音合成评测、系统的可用性和界面的评测、语音通信质量的评测、文字识别系统的评测。

这13项内容的研究对象都是自然语言，当然都涉及语言学。这些研究都要对语言进行形式化的描述，建立合适的算法，并在计算机上实现这些算法，因此，要涉及数学、计算机科学和逻辑学。口语输入、书面语输入、口语输出、信息传输与信息存储都需要电子工程的技术。多模态的计算机处理和话语分析涉及心理学，自然语言系统的评测也需要心理学的理论支持。空间和时间的表征方法涉及哲学，机器词典和词网的建设需要对知识进行分类，需要本体论（ontology）的支持，也涉及哲学。书面语料库和口语语料库的加工需要使用统计方法，涉及统计学。神经网络的连接主义技术涉及生物学。可以看出，从横的角度来考察，自然语言处理也涉及语言学、计算机科学、数学、心理学、哲学、逻辑学、统计学、电子工程、生物学等领域。

不论从纵的角度还是从横的角度来观察，自然语言处理都是一个多边缘的交叉学科。由于它的对象是语言，因此，它基本上是一个语言学科，但它还涉及众多的学科，特别是计算机科学和数学。以上我们从共时方面考察了自然语言处理的学科定位，下面从历时方面来考察这个问题。

3. 自然语言处理的历史

在历史上，自然语言处理曾经在计算机科学、电子工程、语言学和心理认知语言学等不同的领域分别进行研究。

从20世纪40年代到50年代末这个时期是自然语言处理的萌芽期。这类研究最早可以追溯到第二次世界大战刚结束时，那时计算机刚刚发明。在自然语言处理的萌芽期，有两项基础性的研究特别值得注意：一项是Turing算法计算模型的研究，另一项是Shannon概率和信息论模型的研究。

20世纪50年代提出的自动机理论来源于Turing在1936年提出的算法计算模型，这种模型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的工作导致了Kleene关于有限自动机和正则表达式的研究。Turing是一个数学家，他的算法计算模型，与数学有着密切的关系。

1948年，Shannon把离散马尔可夫过程的概率模型应用于描述语言的自动机。1956年，Chomsky从Shannon的工作中汲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，但是，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。

这个时期的另外一项基础性工作是用于语音和语言处理的概率算法的研制，这是Shannon的另一个贡献。Shannon把通过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。这些研究与数学和统计学有着密切的关系。

1946年，König等还研究了声谱。声谱和实验语音学的基础研究为尔后语音识别的研究奠定了基础。这导致了50年代第一个机器语音识别器的研制成功。1952年，Bell实验室的研究人员建立了一个统计系统来识别由一个单独的说话人说出的10个任意的数目字。该系统存储了10个依赖于说话人的模型，它们粗略地代表了数目字的头两个元音的共振峰。Bell实验室的研究人员采用选择与输入具有最高相关系数模式的方法，达到了97～99%的准确率。这些研究与电子工程密切相关。

在20世纪50年代末期到60年代中期，自然语言处理明显地分成两个阵营：一个是符号派（symbolic），一个是随机派（stochastic）。

符号派的工作可分为两个方面。

一方面是50年代后期以及60年代初期和中期Chomsky等的形式语言理论和生成句法研究，很多语言学家和计算机科学家的剖析算法研究，早期的自顶向下和自底向上算法的研究，后期的动态规划的研究。最早的完整的剖析系统是Zelig Harris的“转换与话语分析课题”（Transformation and Discourse Analysis Project，简称TDAP）。这个剖析系统于1958年6月至1959年7月在宾夕法尼亚大学研制成功。这些研究都是语言学家和计算机科学家共同完成的。

另一方面是人工智能的研究。1956年夏天，John McCarthy、Marvin Minsky、Claude Shannon和Nathaniel Rochester等学者组成了一个为期两个月的研究组，讨论关于他们称之为“人工智能”(Artificial Intelligence，简称AI)的问题。尽管有少数的AI研究者着重于研究随机算法和统计算法(包括概率模型和神经网络)，但是大多数的AI研究者着重研究推理和逻辑问题。典型的例子是Newell和Simon关于“逻辑理论家”(Logic Theorist)和“通用问题解答器”(General Problem Solver)的研究工作。早期的自然语言理解系统几乎都是按照这样的观点建立起来的。这些简单的系统把模式匹配和关键词搜索与简单试探的方法结合起来进行推理和自动问答，它们都只能在某个领域内使用。在60年代末期，学者们又研制了更多的形式逻辑系统。AI的研究是计算机科学、哲学、生物学、心理学、语言学密切配合的结果。

随机派主要是一些来自统计学专业和电子学专业的研究人员。在20世纪50年代后期，贝叶斯方法(Bayesian method)开始被应用于解决最优字符识别的问题。1959年，Bledsoe和Browning建立了用于文本识别的贝叶斯系统，该系统使用了一部大词典。计算词典的单词中所观察的字母系列的似然度，把单词中每一个字母的似然度相乘，就可以求出字母系列的似然度来。1964年，Mosteller和Wallace用贝叶斯方法来解决在《联邦主义者》(*The Federalist*)文章中的原作者的分布问题。这些研究与统计学和电子工程密切相关。

20世纪50年代还出现了基于转换语法的第一个人类语言计算机处理的可严格测定的心理模型，并且还出现了第一个联机语料库：布朗美国英语语料库(Brown Corpus)。该语料库包含1 000 000 000单词的语料，样本来自不同文体的500多篇书面文本，涉及的文体有新闻、中篇小说、写实小说、科技文章等。这些语料是布朗大学在1963 - 1964年收集的。美国加州大学的华裔科学家王士元在1976年建立了DOC(Dictionary on Computer)，这是一部联机的汉语方言词典。这些研究成果是语言学和计算机科学相结合的产物。

自然语言处理萌芽期的这些出色的基础性研究，奠定了有关理论和坚实的技术基础。这类研究从一开始就带有明显的交叉学科的特点，它是在各个相关学科的交融和协作中萌芽成长起来的。

20世纪60年代中期到80年代末期是自然语言处理的发展期。在这一期间，各个相关学科彼此协作，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)。这些模型分别由两支队伍独立研制而成。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的Baker等。Baker受到普林斯顿防护分析研究所的Baum及其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。这些都是统计学方法在自然语言处理中应用的成果。

逻辑方法在自然语言处理中取得了很好的成绩。1970年，Colmerauer及其同事们使用逻辑方法研制了Q系统(Q-system)和变形文法(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提出的定子句文法(Definite Clause Grammar)也是在自然语言处理中使用逻辑方法的成功范例之一。1979年Kay对于功能语法的研究，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，这是数学、逻辑学和语言学相结合的可喜收获。

自然语言理解也取得了明显的成绩。这个时期的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积木块。迄今我们还没有看到如此复杂和精妙的系统。这个系统还首次尝试建立基于Halliday系统语法的全面的(在当时看来是全面的)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，耶鲁大学的Roger Schank及其同事和学生(经常被称为耶鲁学派)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。他们的研究多使用基于网络的语义学理论，并且在表达方式中引进Fillmore在1968年提出的关于格角色的概念。这些工作是语言学、计算机科学、数学巧妙结合的成

果。

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。

话语分析 (discourse analysis) 集中探讨了话语研究中的四个关键领域：话语子结构、话语焦点、自动参照消解、基于逻辑的言语行为。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”的框架，即 BDI (Belief-Desire-Intention) 的框架。这样的研究与心理学、逻辑学、哲学有密切关系。

在 1983 - 1993年的十年中，自然语言处理研究者对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，有关的研究又回到了 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”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，显示了这种模型仍然有着强大的功能，从而重新得到自然语言处理界的注意。

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、名词短语附着歧义的判定以及从语音识别到语义学的连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令人瞩目的成绩。

从 20世纪 90年代开始，自然语言处理进入了繁荣期。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，英国著名学者哈钦斯 (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，从此自然语言处理进入了繁荣期。

1994至 1999年，自然语言处理的研究发生了很大的变化，出现了前所未有的局面。这主要表现在三个方面。

首先，概率和数据驱动的方法几乎成为了自然语言处理的标准方法。句法剖析、词类标注、参照消解和话语处理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评测方法。

其次，由于计算机的速度和存储量的增加，使得在语音和语言处理的一些子领域，特别是在语音识别、拼写检查、语法检查这些子领域，有可能进行商品化的开发。语音和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。

最后，网络技术的发展使得基于语言的信息检索和信息抽取的需要变得更加突出。可以预见，网络技术的进一步发展，一定会把自然语言处理的研究推向一个新阶段。

自然语言处理在 50多年的发展历程中，把语言学、计算机科学、数学、心理学、哲学、逻辑学、统计学、电子工程、生物学等学科融合起来，形成了一门边缘性的交叉学科。因此，无论从共时的方面考察，还是从历时的方面考察，我们都可以看出自然语言处理的学科交叉性和边缘性，它横跨了文科 (语言学、哲学、逻辑学)、理科 (计算机科学、数学、心理学、统计学、生物学) 和工科 (电子工程) 三大知识领域，这就是自然语言处理在人类整个知识体系中的定位。

4. 当前自然语言处理发展的特点

21世纪以来，由于国际互联网的普及，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自然语言处理的研究成果来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有关的研究，投入了大量的人力、物力和财力。

当前国外自然语言处理研究有三个显著的特点：

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自然语言处理的主要战略目标：在过去的 40多年中，从事自然语言处理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，其数量之浩瀚和颗粒度之精细，都是以往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、表征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然语言处理问题。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 COLING90)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自然语言处理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 (即 TMI-92)上，所宣布的主题是“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义”，就是指以生成语言学为基础的方法，所谓“经验主义”，就是指以大规模语料库的分析为基础的方法。从中可以看出当前自然语言处理关注的焦点。当前语料库的建设和语料库语言学的崛起，正是自然语言处理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，就必须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，自然语言处理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。传统语言学中啧啧称道的所谓“例不过十不立，反例不过十不破”的朴学精神，貌似严格，实际上，在浩如烟海的有关语言使用的数据中，以十个正例或十个反例就轻率地决定语言规则的取舍，绝不能保证这些规则的可靠性。当前的自然语言处理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，成为了当前这个领域的热点。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，都应该注意到这样的变化，逐渐改变获取语言知识的手段。2003年 7月，在美国马里兰州巴尔的摩，由美国商业部国家标准与技术研究所 (NIST/TIDES: National Institute of Standards and Technology)主持的机器翻译评比中，德国 Aachen大学的博士生奥赫 (Franz Och)获最好成绩，他使用统计方法从双语语料库中自动地获取语言知识，建立统计机器翻译的规则，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。希腊科学家阿基米德说过：“只要给我一个支点，我就可以移动地球。”而现在奥赫也模仿着阿基米德说：“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。”这反映了新一代研究者朝气蓬勃的探索精神和继往开来的豪情壮志。看来，奥赫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，使用机器自动学习的方法，也许有可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新的曙光。过去我们使用人工编制语言规则的方法来研制一个机器翻译系统，往往需要几年的时间，而现在采用奥赫的机器学习方法，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，这是令我们感到振奋的。

第三，自然语言处理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，自然语言处理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。

在这样的新形势下，自然语言处理这个学科的交叉性和边缘性显得更加突出了，我们研究者如果只是局限于自己原有的某个专业领域而不从其他相关的学科汲取营养来丰富自己的知识，在研究中必将一筹莫展、处处碰壁。面对这样的形势，我们绝

不能抱残守缺，继续蜷缩在某个狭窄的领域之内孤芳自赏，而应该与时俱进，迎头赶上，努力学习新的知识，以适应学科交叉性和边缘性的要求。这是我国自然语言处理工作者必须考虑的大问题。

我国的自然语言处理研究虽然已经取得不少成绩，但是与国际水平相比，差距还很大。自然语言处理是国际性的学科，我们不能闭门造车，而应该参与到国际研究中去，用国际水平和国际学术规范来要求我们的研究。近年来，我国的研究人员虽然也到国外参加过第一流的自然语言处理国际会议，如 COLING、ACL、LREC等，但是在这些国际会议上，我国学者从来没有被邀请在会议上做代表当前研究水平的“主题报告”，而是只能在分组会议上讲一讲我们的成绩和体会。这种情况说明，我国的自然语言处理研究，无论在理论上还是在应用系统的开发上，基本上还没有重大的创新。尽管我们的自我感觉良好，但实在还没有什么特别值得称道的突破。我们的研究基本上还是跟踪性的研究，很少有创造性的研究，当然更谈不上具有原创思想的研究了。因此，我们不能夜郎自大，不能坐井观天，我们只有努力学习国外的先进成果，赶上并超过国际先进水平，使我国的自然语言处理在国际先进行列中占有一席之地，才无愧于我国大国的地位。

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.

[2]冯志伟. 应用语言学新论 ——语言应用研究的三大支柱 [M]. 北京，当代世界出版社，2003.

[3]冯志伟. 机器翻译研究 [M]. 北京，中国对外翻译出版公司，2004.

[4] Jurafsky, Daniel & 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.

[5] Kai-Uwe, Carstensen, et al. *Computerlinguistik und Sprachtechnologie, Eine Einführung* [M]. Heidelberg/Berlin: Spektrum Akademischer Verlag, 2004.

[6] Manaris, Bill. Natural language processing: A human-computer interaction perspective [J]. *Advances in Computers* 47 (1999).

(责任编辑 严辰松)

·会讯·

中国英汉语比较研究会第七次全国学术研讨会
征文通知

中国英汉语比较研究会第七次全国学术研讨会定于 2006年 10月在烟台大学召开，现将征文通知公布如下，欢迎学会内外学者就以下议题撰写论文。

一、汉英语言对比研究：1. 对比语言学学科史；2. 对比语言学与语言共性和个性；3. 对比语言学与认知；4. 对比语言学与汉语字本位理论；5. 英汉语篇与互文性；6. 英汉语篇图式；7. 英汉语篇生成机制；8. 英汉语言各层次（语音、词汇、句子、句群、语篇、语义、语用、文体、修辞等）的对比。

二、中西文化比较研究：1. 英汉文化异同；2. 英汉文化与认知；3. 中西文化史；4. 全球化语境中的文化互动；5. 英汉文化比较方法；6. 中西文化价值系统。

三、翻译研究：1. 翻译批评和翻译批评学；2. 翻译伦理；3. 翻译研究范式和方法论；4. 翻译理论史；5. 翻译教学；6. 汉语典籍英译理论；7. 典籍英译译文比较；8. 典籍英译的标准和过程；9. 中国译学建设现状。

论文用中英文撰写均可，请将全文（A4纸，汉字 5号字，行距 1.5）一式 5份于 2006年 4月 10日前寄至：山东省烟台大学外国语学院周国辉教授（邮编 264005，电话 0535-6902006(O)，6902956（H），手机 13791246110），或李晓晖老师（电话 0535-6902716，E-mail: grace_725@163.com）。论文经评审通过后于 2006年 5月发正式开会通知。

中国英汉语比较研究会
烟台大学外国语学院
2005年 5月 7日