

载《国家社会科学基金项目成果选介》，2009年出版

国家社科基金项目《计算语言学方法研究》成果简介

冯志伟 执笔

项目批准号：03BYY019

项目名称：计算语言学方法研究

成果形式：专著

成果字数：498,234字

是否出版：因为有一些问题还不够成熟，还需要进一步研究，作者建议暂不出版

项目负责人：冯志伟（教育部语言文字应用研究所）

课题组成员：杨泉，胡凤国，张和友

计算语言学(computational linguistics)是用计算机研究和处理自然语言的一门新兴边缘学科，涉及语言学、计算机科学、数学、心理学等部门。在计算语言学的发展过程中，提出了很多方法，这些方法，在理论上有一定的深度，在实践上有实用价值，值得引起我们语言学研究者的重视。但是，国内计算语言学界对于这些方法的研究基本上是支离破碎的，缺乏系统的总结，更缺乏理论上的分析。本课题在全面调查国内外计算语言学各种方法的基础上，对这些方法进行了系统的描述，并在理论上进行了深入的分析 and 概括，总结出规律性的具有方法论意义的认识。

其主要内容分为七个部分。

一、计算语言学的学科定位和主要方法

这一部分首先从计算机处理自然语言的过程、计算语言学的范围以及计算语言学的历史三个角度来考察计算语言学的学科定位问题。从计算机处理自然语言的过程来考察它的学科定位，是从纵的角度来讨论；从计算语言学的范围来考察它的学科定位，是从横的角度来讨论。通过这种纵横交错的考察，我们对于计算语言学的学科定位就可以在共时的平面上得到比较清晰的认识。然后，我们再从计算语言学的历史来考察，也就是从发展的角度来讨论，这样，我们对于计算语言学的学科定位就可以在历时的平面上得到比较清晰的认识。

对于计算语言学方法的研究，可以从方法论的角度来论述，也可以从语音、词汇、形态、句法、语义、语用研究中使用的方法来论述。

从方法论的角度，计算语言学方法可以分为基于规则的方法（rule-based approach）和基于统计的方法（statistics-based approach）两个方面。基于规则的方法是理性主义的方法，基于统计的方法是经验主义的方法。这两种方法实际上并不是完全对立的，它们各有利弊，而且目前这两种方法有合流的倾向，它们正在相互结合起来，取长补短，相得益彰。本项目如果把基于规则的方法和基于统计的方法分割开来研究，很多问题将会纠缠不清，不便于论述。因此，本项目不采取这样的论述方式。

本项目采取按照语言学学科分类的方式，从语音、词汇、形态、句法、语义、

语用研究中使用的计算语言学方法来加以论述。分别讨论语音的自动处理方法、词汇的自动处理方法、形态的自动处理方法、句法的自动分析方法、语义的自动处理方法、语用的自动处理方法。

在论述时，首先对于各个领域内计算语言学方法的发展历史进行简要的回顾，然后，再对各种具体的方法进行论述和分析。这样，计算语言学方法的研究便有了一个可靠的历史背景，我们对于各种方法的来龙去脉也就更加清楚了。

二、语音的自动处理方法

文本-语音转换（Text-to-Speech 简称 TTS）的核心任务是以文本中词的序列作为输入，产生声学波形作为输出。自动语音识别（Automatic Speech Recognition, 简称 ASR）的核心任务是以语音的声学波形作为输入，产生单词串作为输出。

这一部分详细讨论了语音自动处理的主要方法：贝叶斯公式（Bayes formula）、噪声信道模型（Noisy Channel Model）、N 元语法（N-gram Grammar）、隐马尔可夫模型（Hidden Markov Model, 简称 HMM）等。这些方法成为了计算语言学中各种统计方法的基础。

在标音方面，本项目采用了美国 DARPA 提出的 ARPABET 代替普通的国际音标 IPA，这种新的标音方法与 ASCII 码一致，便于在计算机上使用，也便于撰写电子文本和印刷排版。

三、词汇的自动处理方法

语言中的词汇具有高度系统化的结构，正是这种结构决定了单词的意义和用法。这种结构包括单词本身的固有的与上下文无关的语义特征以及在文本中单词与单词之间语义关系特征。前者是单词的静态语义特征，后者是单词与单词之间的动态语义特征。

对于单词的静态语义特征，这一部分从知识本体（ontology）的高度出发，分析了美国普林斯顿大学研制的词网（WordNet），指出了其优点和不足之处，并介绍了我国学者提出的 Ontol-MT 通用知识本体系统，说明了 Ontol-MT 在机器翻译和歧义消解中的应用。

对于单词与单词之间的动态语义特征，这一部分介绍了美国语言学家 Fillmore 研制的框架网络（FrameNet）。框架网络的中心思想是词的意义描述必须与语义框架相联系。框架是信仰、实践、制度、想象等概念结构和模式的图解表征，它为一定言语社团中意义的互动提供了基础。

由于多义词是任何语言中都普遍存在的现象，而多义词中诸多的词义分布又很不容易找到一般的规律，多义词的自动排歧涉及到上下文因素、语义因素、语境因素，还涉及到甚至日常生活中的常识，而这些因素的处理，恰恰是计算机最感棘手的问题。所以，词义排歧（Word Sense Disambiguation, 简称 WSD）是计算语言学中的一个特别困难的问题。这一部分分析了英语中的词汇歧义现象，介绍了几种重要的词义排歧方法。

四、形态的自动处理方法

不论是分析型语言、屈折型语言还是黏着型语言，都有形态自动分析的问题。

形态分析主要采用有限状态自动机和有限状态转移网络来进行。这一部分详细地介绍了有限状态自动机和有限状态转移网络的基本原理，通过大量实例来具体地说明自动形态分析的方法。

汉语书面文本是连续的汉字串，单词与单词之间没有空白，因此，汉语形态分析的主要任务就是自动切词和自动词性标注。这一部分还分析了汉语书面文本中确定切词单位的某些形式因素，为自动切词提供了比较可行的方法论基础。

五、句法的自动分析方法

句法自动分析在计算语言学中叫做剖析（parsing）。所谓剖析，就是取一个输入并产生出表示这个输入的结构的过程。所谓句法剖析（syntactic parsing），就是计算机识别一个输入句子并且给这个句子指派一个句法结构（例如，树形图，线图）的过程。

这一部分分别讨论了目前在计算语言学中广泛使用的基于转移网络的自动句法分析方法、基于上下文无关语法的自动句法分析方法、基于特征结构的自动句法分析方法、基于依存语法的自动句法分析方法。

六、语义的自动处理方法

语言的意义可以使用形式化的方法来捕捉，这种形式化方法叫做“意义表示”（meaning representation）。之所以需要这样的意义表示，其原因在于：不论是没有加工过的语言输入，还是用自动句法分析方法推导出来的结构，都不能形式化地表示出语言的意义。因此，这样的“意义表示”能够在从语言输入到与语言输入意义有关的各种各样的具体任务所需要的非语言知识之间架起一座桥梁。我们取语言的输入来构造意义表示，这样的意义表示要使用那些与表示日常生活中的常识性的世界知识同样的材料来构成。产生这样的意义表示并且把它们指派给语言输入的过程叫做“语义分析”（semantic analysis）。

这一部分分别讨论了语言意义的四种表示方法：一阶谓词演算（First Order Predicate Calculus，简称 FOPC）表示法，语义网络（semantic network）表示法，概念依存图（Conceptual Dependency diagram）表示法，基于框架的表示法（Frame-based Representation）。这些意义表示方法都可以把语言输入同外界世界和我们关于外界世界的知识联系起来。

这一部分还讨论了句法驱动的语义自动分析方法、结构语义学、优选语义学、孟塔鸠语法以及意义文本理论。

七、语用的自动处理方法

语用学是对语言与使用环境之间关系的研究。使用环境包括像人和物这样的本体，因此语用学涉及如何将语言用于指示（以及回指）人和物的研究。使用环境也包括话语的上下文，因此语用学也涉及话语结构的形成以及会话时听话人如何理解谈话对象的研究。

语用的自动分析才刚刚开始，国外已经取得初步的成果，国内的研究还做得不多。这一部分主要讨论所指判定和文本连贯的自动分析方法。

本课题的目的在于总结国内外的计算语言学方法，使之系统化，理论化，具体化。由于方法的研究是自然语言处理系统（诸如机器翻译、语料库、信息检索、信息抽取、文本分类等）的开发的关键问题，因此，本课题的研究成果，对于各种类型的自然语言处理实用系统的开发，在方法上具有普遍的指导意义，对于解决我国当前在自然语言信息处理中的理论和现实问题，具有重要的推动作用。

我们已经将本课题中总结出来的一些方法运用于中文信息处理的研究，效果

良好。另外，我们在博士生教学中使用了本课题的部分成果作为教学内容，更新了教材，使学生的眼界为之一新，学生们反映良好。这些都反映了本课题研究成果的应用价值。

由于本课题研究内容很新，涉及面非常广泛，牵涉到语言学、数学、计算机科学很多方面的专门知识，难免有疏失的地方，其中有的问题的考虑还不周全，还有待进一步研究，因此，课题组负责人冯志伟向社科规划办公室建议暂时不出版，也不申请出版资助，目前仅仅作为项目负责人所在高等学校计算语言学专业的硕士生或博士生的教材使用，等条件成熟之后再考虑本成果的出版问题。

作者：冯志伟

职称：研究员，教授，博士生导师

通信地址：100010 北京朝内南小街 51 号

电话：010-6526-7983

手机：15911066112

电子邮箱：zwfengde2008@hotmail.com

“计算语言学方法研究” 社科基金课题评审意见综合

在信息社会中，语言信息处理已成为国家科技产业发展的重要领域。中文信息处理技术的发展离不开理论的指导，离不开科学方法的研究和运用。二十世纪下半叶以来，随着语言信息处理技术在社会上日益受到人们的重视，国外关于计算语言学理论和方法的研究有了长足的进步，取得了丰硕的成果。国内在中文信息处理上也做了大量的工作，研制出一些实用化系统，如机器翻译、自动文摘、人机对话、语音识别、语音合成等等，也做出了一些很好的以资源开发为主的语言工程，如一些大规模的文本语料库、语音语料库甚至多模态语料库等等。但毋庸讳言，在计算语言学理论体系的建设和方法的探讨上，国内的研究基本上是处于一种滞后状态，虽也有很多闪光的东西，但只是零金碎玉，不成系统。从这个意义上来说，该项目的研究成果《计算语言学方法研究》具有开拓性、创新性，起到了填补该领域空白的作用。

该项目的研究特色在于：

一，内容全面，融合中外。计算语言学是一门国际性的学科，国外的计算语言学方法研究一直处于领先地位，国内该领域研究也有很多成果。该项目能够广泛搜集、融合国内外计算语言学理论和方法，使之系统化、条理化，尤其在吸收国外最新研究成果方面做得非常到位，该成果堪称是集国内外计算语言学研究方法之大成者；

二，语言为纲，建立体系。语言信息处理，说到底是对语言现象的处理，计算语言学，本质上是语言学。因此，该项目成果在第一章综述计算语言学的学科定位和主要方法以后，

就以语音、词汇、形态、句法、语义、语用的自动处理方法为纲来安排各个章节，这不能不说是匠心独运，新意迭出；

三，注重实践，事实验证。理论联系实际，也是该成果的一个特色。实践是检验理论、检验方法的最好试金石。计算语言学，不仅是一个理论学科，更是一个实验学科、工程学科。该成果在介绍各种计算语言学理论和方法时，努力结合自然语言实际，在对语言事实的分析中验证理论的科学性和方法的有效性，增加了文章的可读性，不仅适合理科背景的人阅读，也适合语言学背景的人学习；

四，引证丰富，用例鲜活，表现出研究者开阔的学术视野和多元的学术背景。该研究不仅涉及语言学、计算机科学，还涉及数学、心理学、统计学、哲学、信息论等多个学科领域，交叉性很强，但研究者旁征博引，显示出了对多种学科领域都驾轻就熟的能力。计算语言学从产生到今天虽然只有半个多世纪的时间，然而由于涉及多种学科，研究领域广泛，已积累了很多理论和方法，有了比较深厚的学术积淀，但系统化、条理化的程度还不够，国内尤其如此。要从人们那些分散、复杂、繁琐的具体研究中把计算语言学方法梳理出来，条分缕析，形成系统，是一项非常艰难、非常辛苦的工作，作者为之付出了大量的心血和劳动，其精神可敬可嘉。

这样一部计算语言学研究方法的集大成式作品，有着相当高的学术价值和应用价值。

首先，它填补了国内这方面研究的空白，使得国内在计算语言学理论和方法的系统化研究上前进了一大步；其次，由于有语言事实分析做支撑，这些理论和方法可以给在语言信息处理一线的工作人员以具体指导；第三，该著作语言流畅，通俗易懂，可以作为高校里计算语言学专业学生的教材和参考教材，在培养人才上发挥重要的作用。

该研究的不足之处是有些领域中近年来人们使用的新方法还有遗漏。如在汉语自动分词方面，解决交集型歧义切分的“查表法”、“排序法”，解决组合型歧义切分的“词典规则互动法”以及最新的基于统计的“由字构词法”等都没有涉及。虽然所有领域使用方法都全面介绍不可能，但自动分词毕竟是汉语信息处理特有的现象，也是汉语信息处理的基础，不容忽视。另外，作为理论性著作，篇幅显得有些长，有些论述和举例还可简省些。