

【当代语言学论丛】
[论文集]

语言应用研究(第二集)

亢世勇 主 编 YU YAN YING YONG YAN JIU

中国文史出版社

山东省“十五”期间重点学科强化建设项目

鲁东大学汉语言文字学科语言文字理论与应用研究文库(二)

语言应用研究

(第二集)

亢世勇 主编

中国文史出版社

图书在版编目(CIP)数据

语言应用研究(第二集)/亢世勇主编. - 北京:中国文史出版社,2006.8

(当代语言学论丛/陈淑梅,卢小林主编)

ISBN 7-5034-1822-2

I. 语… II. 亢… III. 汉学—应用语言学—研究 IV. H1

中国版本图书馆 CIP 数据核字(2006)第 068517 号

语言应用研究(第二集)

责任编辑:李春华 封面设计:福瑞来书装

出版发行:中国文史出版社

社 址:北京太平桥大街 23 号 邮编:100811

印 刷:北京凯通印刷厂

经 销:新华书店北京发行所

开 本:880×1230 毫米 1/32

印 张:10.5

字 数:292 千字

版 次:2006 年 8 月北京第 1 版

印 次:2006 年 8 月第 1 次印刷

全套定价:150.00 元

文史版图书如有印、装错误,工厂负责退换。

总 序

张志毅

乙酉鸡年冬至后三日,正是“气始于冬至,周而复生”的日子,我们的省级强化建设学科汉语言文字学又传来喜讯:新的一批研究成果,作为语言文字理论与应用研究文库,即将出版,囑写一序。于是便目睹了大饱眼福、大饱口福的方丈盈前:

张绍麒教授主编的《汉外词汇对比研究报告》,

陈淑梅教授主编的《词汇语义学论集》,

亢世勇教授主编的《语言应用研究(第二集)》,

徐德宽博士的《信息时代的语言教学与研究》,

王宝刚博士的《〈方言〉简注》,

解海江博士的《汉语词汇对比研究探索》,

李海英副教授的《普通话水平测试的社会语言学阐释》,

姜 岚副教授的《威海方言调查研究》,

姜仁涛讲师的《〈尔雅〉同义词研究》,

张文峰、侯仁魁老师的《计算语言学》。

这真是珍肴异饌。孰能与之媲美呢?

那就是我从年初到年末看的俄语、汉语、英语多篇应届博士论文。比外语博士论文,我们的成果少了点匠气,多了点师魂。在引介和运用外国新理论方面,我们稍逊于人;在基于语料库,脚踏实地的升华方面,我们略胜于。在几个博士点座谈中,我多次强调“脱去匠气,生发师魂”。

到了今日的地球村,纯国粹的课题已经寥若晨星。因此对绝大多

数的专题都必须极力扩展视野,扩展到古今中外,每遇一题必须梳理其中外学术思想史。否则,谁都难免坐井观天,连王安石这样的泰斗对国粹的“十三经”也偶尔“寡识不知周礼伪”,而我们这些初出茅庐的新手更要小心翼翼地观测天高地厚。只有理清中外学术思想史,才能认清某一说法的新旧、前后、高低。否则,可能扮演了悲剧角色(以“旧”为“新”)而自以为新,这在许多专著和教材中时有发生。学术要堕落到这等地步,那也真是泡沫了。好在“阳乘阴,是以万物仰而生”,一股朴学而清新的学术空气,正像冬至阳气又昂然升起,学术百花园又将争奇斗艳。以是记于盈前方丈。

2005 年圣诞节

目 录

自然语言处理的学科定位(代序)	冯志伟(1)
现代汉语系统语料库的建设与相关研究	亢世勇(22)
词汇语义学与知识本体	冯志伟(38)
Ontology 及在自然语言处理中的应用	裴亚军(77)
基于数据库的汉语双音合成词语素义与义位关系研究	许小星(85)
汉语双音合成词语义构词特点及原因分析	韩 晓(102)
从“字本位”看《说文解字》“人部”的字义网络系统	孙道功(114)
英汉新词语理据对比分析	李松芬(125)
新数字词研究	徐小波(137)
新词词群中的中心词素浅析	张青琳(150)
《新词语大词典》中多音节新词语的特点	宋宏灿等(158)
网络用语“谐音”现象探析	樊立三(166)
网络汉字词“旧词赋新义”现象初探	李云云(176)
汉语外来词多译并存现象初探	宋 华(187)
句子语义成分标注规范	李毅等(196)
基于标注语料库的现代汉语单句句型句模的对应关系研究	孙道功等(207)
现代汉语复句形式的句模	田珍都等(222)

现代汉语语文辞书词性标注现状分析	樊立三等(234)
基于语料库的指示代词“那个”“那些”语法特征分析	李娜(252)
基于语料库的旁指代词“其他”“其它”“其余”的语法特征考察	韩晓等(268)
基于数据库的成语中的数字英译初探	李衍妮(276)
基于语料库的汉英数词谚语翻译分析	孙慧明(290)
中级阶段留学生语篇感的培养	孙雁雁(304)
莒县话儿化音节的结构	钱永文(312)
后 记	(322)

自然语言处理的学科定位(代序)

冯志伟

教育部语言文字应用研究所

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,五十多年来,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,使我们认识到自然语言处理在学科体系中的位置,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方面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方面得到比较清晰的认识。

1. 从自然语言处理的过程来考察

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。

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。

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:

(1)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。

(2)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。

(3)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。

(4)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。

(5)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。

(6)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。

(7)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。

(8)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涵义。

(9)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的涵义的知识。

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涵义,并且执行这个指令,一般来说需要经过如下的处理过程:

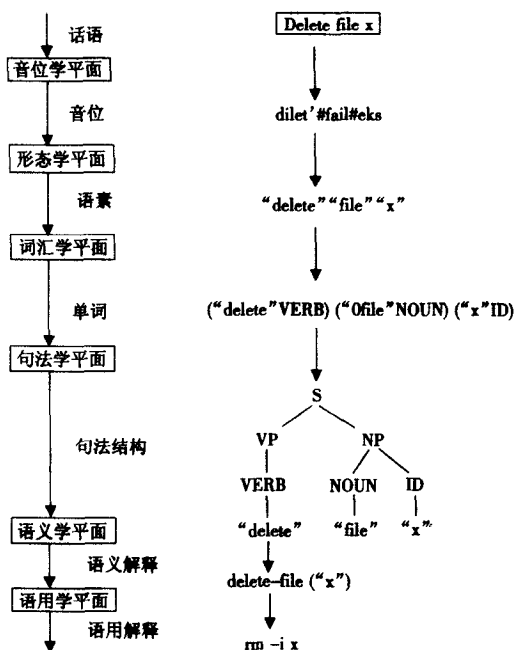


图 自然语言处理系统中的知识平面

从图中可以看出,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 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,上述这些语言学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到如下的知识领域:

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。

逻辑学:给自然语言处理提供逻辑运算和逻辑推理的理论和方法。

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。

由此可见,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,一个人很难对于上述各个领域的知识门门皆通,但是,至少对于他自己的专业领域应该是精研通达的内行,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. 从自然语言处理的范围来考察

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到众多的部门,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, 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十三项内容:

(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(Handwriting as Computer Interface):例如,用笔输入的计算机(pen computer)

手写文字分析(Handwriting Analysis):例如,签名验证(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(Grammar Formalisms):例如,上下文无关语法(context Free Grammar, 简称 CFG),词汇功能语法(Lexical Functional Grammar, 简称 LFG),功能合一语法(Function Unification Grammar, 简称 FUG),中心语驱动的短语结构语法(Head - Driven Phrase Structure Grammar, 简称 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(Computational 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(Summarization):例如,自动文摘(ext 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)]

文本与图象处理(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 Inte-

gration: 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(Connectionist Techniques)例如,神经网络(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)

术语编纂与术语数据库(Terminology and Terminological Databank)

网络数据挖掘与信息提取(data - mining and information extract in Web)

(13) 自然语言处理系统的评测(Evaluation)

· 面向任务的文本分析评测(Task - Oriented Text Analysis Evaluation)

机器翻译系统和翻译工具的评测(Evaluation of Machine Translation and Translation Tools)

大覆盖面的自然语言剖析器的评测(Evaluation of Broad - Coverage