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）

亢世勇摇主编

中国文史出版社

摇摇

图书在版编目（悦陨孕）数据

语言应用研究（第二集）辕亢世勇主编援原北京 颐中国文史出版社，圆园园缘援

（当代语言学论丛 辕陈淑梅，卢小林主编）

陨杂月晕 苑原缘园猿源原员远愿圆原员原圆

Ⅰ援语… Ⅱ援亢… Ⅲ援汉学—应用语言学—研究 Ⅳ援匀员

中国版本图书馆 悦陨孕 数据核字（圆园园缘）第 园园愿怨苑远 号

语言应用研究（第二集）

责任编辑：李春华摇封面设计：福瑞来书装

出版发行：中国文史出版社

社摇摇址：北京太平桥大街 圆猿 号摇摇摇邮编：员园园愿员员

印摇摇刷：北京凯通印刷厂

经摇摇销：新华书店北京发行所

开摇摇本：愿缘园×员员远愿 毫米摇员辕猿圆

印摇摇张：远园援缘

字摇摇数：远圆园 千字

版摇摇次：圆园园缘 年 愿月北京第 员版

印摇摇次：圆园园缘 年 愿月第 员次印刷

全套定价：远愿园援园园 元

文史版图书如有印、装错误，工厂负责退换。

总摇摇序

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，嘱写一序。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《约方言 跃简注》，

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《约尔雅 跃同义词研究》，

张文峰、侯仁魁老师的《计算语言学》。

这真是珍肴异馔。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，我们略胜于人。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，纯国粹的课题已经廖若晨星。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。

圆园园缘年圣诞节

目摇摇录

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

引摇言 :采用计算机技术来研究和处理自然语言是 圆园世纪 源园年代末期和 缘园年代才开始的,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 ([illegible] 简称 晕蕴孕。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。

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 [illegible] 在 [illegible] 年出版的《计算机进展》([illegible] 第 [illegible] 卷的《从人 [illegible] 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([illegible] 和语言应用([illegible] 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:"[illegible]

[illegible]

[illegible]

[illegible]

[illegible]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。

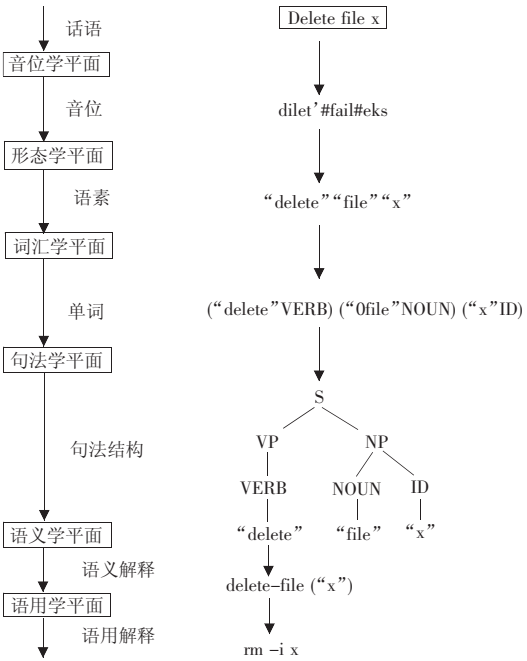
(8)语用学的知识 描述与情景有关的情景语义,说明怎样推导出句子具有的与周围话语有关的各种涵义。

(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:"阅[illegible]"("删除文件载"),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:



图摇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“阅[illegible]”在音位学平面转化成音位系列“[illegible]”,然后在形态学平面把这个音位系列转化为语素系列“[illegible]”“[illegible]”“曾”,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:(“[illegible]”[illegible]),(“[illegible]”[illegible]),(“曾”[illegible]),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:“[illegible]”(“曾”),在语用学平面得到这个指令的语用解释“[illegible]”,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[illegible]为[illegible]设计的一个语音理解界面,叫做[illegible]这个语音理解界面使用了上述的第[illegible]至第[illegible]个平面的知识,得到口头指令“阅[illegible]”的语义解释:“[illegible]”(“曾”),然后,使用第[illegible]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“[illegible]”,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[illegible]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[illegible]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[illegible]至第[illegible]个平面的知识,个别的机器翻译系统还需要第[illegible]个方面的知识;语音识别系统只需要第[illegible]至第[illegible]个平面的知识。

上述[illegible]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。

圆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 (简称 ”,或者称为“语言工程”(

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

(员 口语输入 (

语音识别 (

信号表示 (语音信号分析) [

鲁棒的语音识别 (

语音识别中的隐马尔可夫模型方法 [

语言表示理论 (语言模型) [

说话人识别 (

口语理解 (

(圆 书面语输入 (

文献格式识别 [

光学字符识别 :印刷体识别 [

光学字符识别:手写体识别

手写界面(书写输入):例如,用笔输入的计算机(或称笔输入计算机)

手写文字分析(书写输入):例如,签名验证(或称笔输入验证)

(3)语言分析和理解(或称自然语言理解)

小于句子单位的处理(形态分析,形态排歧)[或称词法分析(广义)和词法排歧(广义)]

语法的形式化(或称语言学家可以理解的语法):例如,上下文无关语法(或称短语结构语法)简称 CFG,词汇功能语法(或称词汇功能语法)简称 LFG,功能合一语法(或称功能合一语法)简称 UG,中心语驱动的短语结构语法(或称中心语驱动的短语结构语法)简称 HPSG

针对基于约束的语法编写的词表(或称词汇知识表示)

计算语义学

句子建模与剖析技术

鲁棒的剖析技术

(4)语言生成(或称自然语言生成)

句法生成

深层生成

(5)口语输出技术

合成语音生成

用于文本-语音合成(或称从文本到语音)的文本解释[或称文本到语音(从概念到语音)]

口语生成(从概念到语音)[或称文本到语音(从概念到语音)]

(6)话语分析与对话

话语建模 (

对话建模 (

口语对话系统 (

(

文献自动处理 (

文献检索 (

文本解释 :信息抽取 (

文本内容的自动归纳 (

:例如,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 (

(

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 (

(

多模态的计算机处理 (

空间和时间的表示方法 (从文本中抽取空间和时间的信息) [

文本与图象处理 (

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 (

[illegible]

人的因素与用户的可接受性 (与[illegible]

语音识别 :评估与评测 ([illegible]

语音合成评测 ([illegible]

系统的可用性和界面的评测 ([illegible]

语音通信质量的评测 ([illegible]

文字识别系统的评测 ([illegible]

这十三项内容的研究对象都是自然语言,当然都涉及到语言学,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,要涉及到数学、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到心理学,自然语言系统的评测也需要心理学的理论支持。空间和时间的表示方法涉及到哲学,机器词典和词网的建设需要对知识进行分类,需要本体论 ([illegible] 的支持,也涉及到哲学。书面语料库和口语语料库的加工需要使用统计方法,涉及到统计学。神经网络的连接主义技术涉及到生物学。可以看出,从横的角度来考察,自然语言处理也涉及到语言学、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,由于自然语言处理的对象是语言,因此,它基本上是一个语言学科,但是,它还涉及到众多的学科,特别是涉及到计算机科学和数学。前面我们从共时方面考察了自然语言处理的学科定位,下面我们进一步从历时方面来考察这个问题。

猿援从自然语言处理的历史来考察

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。之所以出现这种情况,是由于自然语言处理包括了一系列性质不同而又彼此交叉的学科,因