基于语用的自然语言处理研究与应用初探

李 蕾,周延泉,钟义信
(北京邮电大学 智能科学技术研究中心,北京 100876)

摘 要:首先分析了语用信息的必要性和重要性,认为只有融入语用研究的自然语言处理技术才能显示"以人为本"和智能化的特色,只有语用、语义和语法信息的研究都成熟了,才能使计算机真正获得自然语言所表达的信息,达到与人类交流对话的水平.接着介绍了语用学的产生、发展和运用状况,剖析了存在的主要问题,提出了基于语用的自然语言处理.然后结合典型应用背景——奥运多语言信息服务示范终端"City Guide"语音识别后文本的检错纠错需求,探索并尝试了一种基于语用信息的自然语言处理检错纠错方法,并通过真实语料的测试来检验效果.结果表明,当前算法可以使中文语音识别正确率提高 29 %.

关键词:自然语言处理;语用信息;语音识别检错纠错
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-4785(2006)02-0001-06

Pragmatic Information Based NLP Research and Application

LI Lei, ZHOU Yan-quan, ZHONG Yi-xin
(Center for Intelligence Science and Technology Research, Beijing University of Posts & Telecommunications, Beijing 100876, China)

Abstract: Pragmatic information is looked on as the next focus for natural language processing (NLP) research. The necessity and importance of pragmatic information are analyzed firstly. It is pointed out that NLP could be charaterized as humanity and intelligence only after pragmatic information are integrated into it. And only when syntactic, semantic and pragmatic information are all fully studied could computers understand the information expressed in human natural language. Thus computers could really communicate with human. Then details of pragmatics research are introduced, including its origin, growing history and applications. Problems are also analyzed for its current status. As a result, pragmatic information based NLP is put forward. Then a grope research of this, i.e. the sentence error detection and correction in the application domain of "City Guide" Speech Recognition (SR) interface is reported. The "City Guide" is a demo terminal for the National 863 project of "Olympics Oriented Multilingual Information Service". A method containing pragmatic information analysis is studied and tested using realistic corpus. Results show that the precision of Chinese SR can be improved by 29 %.

Keywords: natural language processing (NLP); pragmatic information; error detection and correction for SR

科学发展到今天,信息技术已经成为无处不在的主流,而其中最能显示"以人为本"特色的,就是自然语言处理技术.认知心理学研究表明,人类的自然语言包含了非常复杂的心理活动,同时也是知觉、记忆、思维等许多不同心理活动的主要参与者.事实上,自然语言是一个复杂的系统,具有层次结构[1].作为人类用来表达信息的工具,自然语言包括形式结构所表达的语法信息、形式结构所包含的逻辑内容所表达的语义信息、形式结构及其包含的逻辑内容一起所体现的、对于特定目的而言的语用信息.可见,只有融入语用信息研究的自然语言处理技术才

收稿日期:2006-05-16.
基金项目:国家自然科学基金资助项目(60575034);国家"863"资助项目(2004AA117010,2005AA117010).

能最终实现科学技术人性化和智能化的目标.

国内外关于自然语言处理与理解方法的研究,长期专注于"语法"层次的研究;20世纪末期以来,人们认识到单纯在语法层次上的研究不能解决问题,开始进到了"语义"的层次,最好的例子就是互联网络这几年的研究正在从WWW走向语义网Semantic Web. 然而,自然语言是语法、语义、语用三者的"有机统一体",只从语法语义2个层次上研究也不能很满意地解决问题. 语用研究的空缺已经日益阻碍了自然语言处理的发展. 只有语用、语义和语法信息的研究都成熟了,才能真正通过分析获得自然语言所表达的信息,达到与人类交流对话的水平. 因此认为,未来的趋势是要充分研究和利用自然语言的语法、语义和语用信息. 实际上,走向语法-语义-语用三位一体的层次是必然的出路,语用作为自然语言中体现效用价值的因素不可能因为困难总被回避,现在已经到了必须要面对的时候了.

1 语用学研究分析

1.1 语用学的由来和两大流派

语用学研究始于语言哲学,其发展一直不被人们重视. 但语言学的历史证明,语言学作为一门科学,与社会有千丝万缕的联系,与人们的生活息息相关,因而人们研究这方面的兴趣也就越来越浓厚. 20世纪60年代提出的"语言学是以人和社会作为实践对象的科学"的观点得到肯定,语用学也就被认为是继承和推动这种观点的重要力量而获得发展. 直到1977年Journal of Pragmatics[2]创刊,以及1986年国际语用学学会(the international pragmatics association,简称IPrA)的创立,标志着语用学学科地位的确立.

当今语用学多指语言学的语用学,主要分为两大流派:其一是英美学派,将语用学看成是语言学的一个分支,称为微观语用学,或是分相论;其二是欧洲大陆学派,主张凡与语言的理解和使用有关的都是语用学的研究对象,将语用学看成是语言功能的一种综观,故称宏观语用学,或是综观论. 分相论一直是语用学界普遍接受的看法,认为指示语、前提、会话含意、言语行为、会话结构等是语用学的基本研究内容. 而综观论则认为语言使用的过程实际上是为了顺应而不断做出语言选择的过程,在所有层面上都值得进行语用研究[3].

1.2 语用学理论的运用

语用学理论在外语教学、对外汉语教学、翻译研究、修辞学研究、汉语研究等领域得到了广泛运用. 如在汉外翻译研究中,翻译的语用等值问题首先成为研究热点,其后"西译汉化"和"汉译西化"问题也引起了热烈的讨论,语用学不断被推向翻译研究的应用之中. 而在修辞学领域,修辞学与语用学之间的关系问题,以及二者在成功的会话交际中所起的重要作用等成为研究热点,为语用学和修辞学提供了新的研究和范围.

在应用中产生了很多语用学的研究分支,如研究语言本身语用问题的语用语言学;研究语言和心理认知关系的认知语用学;研究语用与文化关系的跨文化语用学等. 目前一个新的领域正在引起人们的注意,也是作者多年来思考和关注的一个问题,这就是语用学同形式语言学相结合的形式语用学,它研究语用的形式化,探讨语用学理论在人工智能和计算机自然语言处理方面的作用[3-5].

我国汉语界的突出成就是对语境的研究. 1991年在山东大学召开了第二届全国语用学研讨会,1992年出版了第一部语境研究论文专集《语境研究论文集》. 此后在这个领域也作了很多研究,如区分语言与语境、修辞中零度与偏离的关系、抽象语义与语境语义的关系、言内歧义与言外歧义的关系以及语境对对话理解的作用等[6-7].

1.3 基于全信息(语法、语义和语用信息)的自然语言处理方法的研究

语用学的蓬勃兴起显示出其光明的发展前景,但目前对语用的研究存在局限性,突出表现在:

1) 研究的领域多集中在语言学界,甚至包括哲学界,针对自然语言处理的研究开展的很少. 由于语用分析难度大,几乎很少有研究者将其应用在自然语言处理领域,从而造成了当前自然语言处理中语用研究的空白.

2) 对语言中的某些语用现象进行了较为深入的探讨,但没有形成系统的研究成果,不能为计算机处理自然能语言的迫切需要提供足够的语用知识,这最终也必将影响语言学的健康发展.

基于此,北京邮电大学钟义信教授认为,将全信息,即语用、语法、语义信息应用于信息科学是非常重要的,之后进一步提出将全信息应用于自然语言理解,提出了"全信息自然语言理解方法论",前瞻性地指出和强调了自然语言处理中语用研究的重要

性[8].近年来,北邮智能科学技术研究中心在这方面进行了一系列面向实际应用的研究与开发,如基于理解的中文自动文摘、基于倾向判断的垃圾邮件过滤、基于全信息的中文信息抽取等,取得了一定的效果[9-11].其中都有对语用信息的探讨,用到的方法有效用度空间、效用规则、文本分类等.

2　基于语用的自然语言处理

随着研究的展开和深入,人们越来越体会到语用信息对于自然语言处理系统的重要作用.如果能够将语用信息应用自如,必将克服很多现有系统无法解决的难题,如歧义问题、倾向性问题、可信性问题、有效性问题、领域可移植性问题等.但是相比于语法、语义信息的研究,语用信息的研究基础太薄弱了,已经成为影响自然语言处理系统性能提高的瓶颈.因此,对语用信息展开深入、细致、全面、基础的研究是当前十分必要、重要且紧急的任务.

基于语用的自然语言处理的最大特色,在于它能够模拟人类利用语用信息来解决问题.以搜索为例,在给定环境下搜索到达目标的途径时,对于各种可能途径不再是盲目的选择或系统地探索,而是先估计不同途径对于到达目标而言的效用度,在比较效用度大小的基础上,选择最有希望的途径.因此,盲目性比较小,成功的把握比较大.当然,获得语用信息往往要付出一定的代价,但从发展的观点来看,随着技术本身的不断进步,代价将会越来越小,而由此所带来的得益却会越来越多.因此,从长远来看,充分利用语用信息是一个应当追求的目标.实际上可以认为这是自然语言处理理论发展的一个重要途径,也是一个重要的方向.但是也要注意:在机器上实现的语用和语言学界甚至哲学界所研究的语用不同,不能完全陷入哲学界和语言学界关于"语用"问题的争论中.

对自然语言处理中的语用信息展开深入的研究,至少应当包括语用信息的确定、描述、度量、分析使用方法及其与语法信息和语义信息的关系等.简单地说,自然语言处理中的语用信息就是指自然语言所携带的、针对自然语言处理系统的应用目标而言的信息.而自然语言处理系统的应用目标是多种多样的,由此语用信息也不能一概而论,必须要找到分析和确定各种语用信息的基本原则和基本方法,进一步找到确定语用信息相关因素以及多种因素如何协调统一的基本方法.在确定了自然语言处理系统的语用信息基础上,进一步处理的条件是要能够把语用信息用较好的方式描述出来.因此如何找到好的描述方式也是一个关键问题.而为了实现对语用信息方便高效的计算处理,极有必要找到语用信息的数值度量方法.基于这些,还需要进一步寻找和总结在自然语言处理系统中语用信息的分析使用方法.

此外,语用信息与语法、语义信息的互动关系是否和谐,也是影响自然语言处理系统整体性能的关键问题.语用信息不是空中楼阁,它与语法、语义信息是密切联系在一起的,换句话说,语用信息是在语法、语义信息基础之上才能存在、描述和使用的.因此,明确他们之间的关系和互动方式,充分发挥整体大于部分和的作用,才能最终实现对于自然语言的理解.

文中结合一个典型应用背景,即国家"863"项目"奥运多语言综合信息服务"的典型示范系统"City Guide",对基于语用的自然语言信息处理进行了初步的研究和探索.在分析处理语法信息和语义信息的基础上,研究实现了一种语用信息主导的语音识别后语句检错纠错方法."City Guide"是在智能手机平台上实现的一个信息服务终端,支持语音输入/输出,可为奥运期间来北京的参观旅游者提供住宿、交通、旅游等方面的多语言信息服务.目前该演示系统主要支持单句语音输入,如"今天晚上还有房间吗?"、"请问最近的车站在哪?".但是测试过程中的语音输入识别效果很差,达不到实用化的要求.初步实验结果表明,增加了语用分析的自然语言理解可以在一定程度上提高语音识别正确率,关键问题就是如何更好的挖掘和使用语用信息.

3　语用信息主导的语音识别后语句检错纠错

"City Guide"为了克服移动终端屏幕小、使用不方便的问题,采用了语音人机交互方式.语音方式具有自然、方便、快速的特点,目前支持语音功能的人机对话系统已经成为科研和产业界关注的重点.但是语音识别引擎的正确性比较低,如何才能有效地提高和确保人机对话的正确性和有效性就成为主要的问题.对此,文中认为语用信息的作用是必不可少的,只有当机器充分理解了用户的对话目的和对话内容,才能克服语音识别引擎的错误,保证对话的正确顺利进行.

传统语音识别的方法无论是基于统计的模型还是基于规则的模型,主要是针对音节信号进行处理和识别,对识别的内容并不进行正确性分析.如用户输入"叫辆出租车",结果显示却可能出现"较量出租车",由于"叫辆"和"较量"在发音上有一定的相似性,采用语音识别的方法很难做出正确的判断,但是如果换一种处理思路,采用自然语言理解的方法分析结果内容,则很容易判断出"较量出租车"不符合人们的用法,是一个错误的结果.这个例子显示了人们对人机对话系统认识的一个误区,人们常常认为,系统的关键技术是语音识别、语音合成、机器翻译等,但通过例子不难看出,问题的难度最终不在语音表层结构的识别-合成,而在它的核心-明确语用的自然语言理解.

文中在语音识别引擎以后引入一个自然语言理解模块,综合语法、语义和语用信息对语音识别结果进行分析、检错和纠错.语法信息方面,主要分析了功能性词语(如祈使性词语、疑问词语等)在语句中的习惯性位置以及不同发音的词语和词语组合在语音识别过程中的稳定程度.语义信息方面,根据"City Guide"系统功能把语句含义分成了9类:饭馆就餐、购买衣服、讨价还价、旅馆住宿、问路、修理、打车、就医、寻人/物.并且为每一类语句建立了一个初始的核心词列表.所谓核心词,是指某一类语句中出现的能够揭示其含义的关键词.与其相对应的普通词则是指在各个类别的语句中都可能出现,不太影响类别含义的词语.

那么人机对话过程中的语用信息是什么?从信息服务提供和信息获取的角度看,人机对话的目的是为用户提供方便的信息和服务,这种服务通常是面向某些特定领域的,因此语用信息要具体分析到所提供信息服务的领域内容,详细考察应用领域的特点,用户使用这些信息服务的需求是如何表达的,用户使用服务时的环境如何等,而这些都是与语法信息和语义信息密不可分的.前面提到的语用信息分析方法中,效用度空间方法比较抽象,效用规则方法难以处理多变的语音识别错误,而文本分类方法线条比较粗,不能很好的满足需要.为此,本文的语用分析算法是以语义类作为基础目标进行的,主要考察在某种应用目的下,当前语句中一个词语与周围环境的协调适应能力.事实上,心理学研究已经证实上下文对句子理解有重大作用[1].作为知识库,本文面向特定领域的应用场景,建立了常识标准下的语用描述库,主要考虑以下情况:

1)一个语句范围内,核心词与核心词之间的协调能力.

2)一个语句范围内,核心词与必要的普通词之间的协调能力.

一期课题从最简单的二维环境入手,采用统计方法考察2个词语的协调能力(定义如下),更高维数的环境还可以基于二维环境来扩展.

$$Harmoney(word_i, word_j) \mid Goal_k = \frac{co\text{-}occurence(Goal_k, word_i, word_j)}{co\text{-}occurence(word_i, word_j)} \times$$

$$\frac{co\text{-}occurence(word_i, word_j)}{occurence(word_i) + occurence(word_j) - co\text{-}occurence(word_i, word_j)} =$$

$$\frac{co\text{-}occurence(Goal_k, word_i, word_j)}{occurence(word_i) + occurence(word_j) - co\text{-}occurence(word_i, word_j)}.$$

其中分开写的2个因子中,第2个因子考察2个词语出现在同一语句中的频度,第一个因子考察2个词语都出现且共同揭示某个语义类的能力,$word_i$、$word_j$ 表示不同的2个词语,$Goal_k$ 表示某语义目标,$co\text{-}occurence(word_i, word_j$ 表示两词语同时出现的所有语句数目,$co\text{-}occurence(Goal_k, word_i, word_j)$ 表示两词语同时出现、并且揭示该语义目标的语句数目,$occurence(word_i)$ 和 $occurence(word_i)$ 分别表示单个词语出现的语句数目.

可见,语用信息必须在语法信息和语义信息的基础上才能分析和表达.语用知识库可由统计方法自动获得,需要收集训练语料.在一期课题的语料基础上,还可以借助公用搜索引擎的帮助,自动获得更多典型语料库.

语句的检错纠错算法就是综合语法、语义和语用信息对各个词语进行可信度评估,如式(1)所示:

$$Reliability(word_i) = Syntactic(word_i) + Semantic(word_i) + Pragmatic(word_i). \quad (1)$$

式中:$word_i$ 是语句中第 i 个词语,该语句中共有 $|s|$ 个词语,$Reliability(word_i)$ 是第 i 个词语的可信

度,包括对第 i 个词语的全信息的考察. $Syntactic(word_i)$ 表示语法信息度量,$Semantic(word_i)$ 表示语义信息度量,$Semantic(word_i)$ 表示语用信息度量. 本期课题中的综合加权采用"等权"规则,还比较简单,今后还需要对各个部分的作用和加权进行深入研究.

通过比较发现一个语句中可信度最小的词汇就是被认为出错概率最高的词汇,也就是纠错的重点. 纠错的基础一方面是找到可能出错的点,另一方面是找到确信正确的点,二者结合才有可能做到较好的纠错. 因此,对于语句中可信度很高的词语也要加以充分利用. 根据词语可信度的度量可以大体确定一个语句中可能正确的点. 根据确信点所需要的全信息环境来修正语音识别结果. 修正过程中可能存在歧义现象,目前的解决方法主要是通过计算词语的汉字内码距离来确定最为接近的纠正结果,因为一方面本期课题中无法直接获得词语音节特征,另一方面汉字内码的编排确是遵循一定的音节顺序的.

显然文中算法是面向特定的应用场景实现的,全信息知识库也都是根据特定应用目的建立的. 因此,实用化存在的一个最大的问题就是领域可移植性. 为此,文中设计实现了一个学习与训练模块,能够根据用户所提供的特定领域语料进行自学习. 如果再有后续的人工参与,就能生成高质量的领域知识库,从而可以方便地完成领域移植.

4　测　试

文中算法采用 VC 编程在 PC 机环境下实现. 为了确定该方法对语音识别后文本的正确率提高是否有作用,采取大规模真实语音识别后文本作为测试用例,编制专门的测试程序,测试了纠错效果.

测试环境:

硬件:

计算机主频 P4 2.0GHz,内存 512MB,硬盘 80GB

操作系统: WindowsXP + SP2

开发平台: VC6.0

语音识别:IBM Viavoice 中文版(自由下载)

语音录入环境:普通办公室工作环境,有背景噪音(电脑、空调、人员对话、走路等),也有突发的干扰,如电话来电的声音、门开关的声音等.

测试结果如表 1 所示.

表 1　测试结果

Table 1　Testing results

项　　目	数值
总测试句子数	1 872
(人工判定)语音识别正确的句子数	663
(人工判定)语音识别错误的句子数	1 209
语音识别正确率	35.4%
(本模块判定)语音识别错误的句子数	959
本模块纠正正确的句子数	544
纠正正确率	56.7%
语音识别 + 本模块正确的句子数	1 207
语音识别 + 本模块正确率	64.4%
比未使用本模块时的正确率提高	29%

由于目前实现的仅是单句功能,所以语句越长,可利用的全信息元素就越多,对语句的处理就可能越深入;反之,语句越短,能利用的信息很少,几乎连其目的都很难确定,这样的错误很难纠正,也是实验中造成纠错失败的重要原因. 目前克服这个问题只能依靠声音距离. 长远的解决方法应该是超越单句功能,以连续对话的信息服务方式进行人机交互,这样便可提供充足的上下文环境,保证机器更好的理解人的需求.

5　结束语

文中在追溯语用学发展历史、剖析其现状的基础上,分析了基于语用的自然语言处理研究的重要性和必要性,认为它必将是下一步自然语言处理的研究重点,并对其主要研究内容进行了初步讨论. 进一步面向"City Guide"应用领域做了具体的研究,设计实现了综合分析语法、语义和语用信息的自然语言处理的语句检错纠错算法,其中重点增加了语用信息的表达和分析. 测试结果表明,该算法可以在一定程度上提高语音识别的正确率. 作为起点,文中的工作给人以足够的信息和勇气,人们将会继续对语用信息展开深入扎实的研究,也期待有更多的研究者加入.

参考文献:

[1]王 [illegible],汪安圣. 认知心理学[M]. 北京:北京大学出版社,1997.

[2]Journal of Pragamatics [EB/OL]. http://www.sciencedirect.com/science/journal/03782166.

[3]何自然,吴亚新. 语用学概略 [EB/OL]. 中国语言学研究会 http://www.pragmaticschina.com,2005.9.

HE Ziran, WU Yaxin. Pragmatics overview[EB/OL]. China Association of Linguistics, http://www.pragmaticschina.com, 2005.9.

[4]刘淑芬. 试论语法分析中句法语义语用的三位一体性[J]. 唐山师范学院学报, 2005,27(6):38-40.
LIU Shufen. A tentative exposition of trinity of syntax, semantics and pragmatics in grammatical analysis[J]. Journal of Tangshan Teachers College, 2005,27(6):38-40.

[5]孙建华. 语境与语用歧义[J]. 河南大学学报(社会科学版). 2004,44(4):142-144.
SUN Jianhua. Context and pragmatic ambiguity[J]. Journal of Henan University (Social Science), 2004,44(4):142-144.

[6]白解红. 语境与语用研究[J]. 湖南师范大学社会科学学报,2000,29(3):88-92.
BAI Jiehong. On context and the use of language[J]. Journal of Social Science of Hunan Normal University, 2000,29(3):88-92.

[7]任万芳. 预设、语境、歧义[J]. 张家口师专学报. 2003. 19(4):24-30.
REN Wanfang. Presupposition, context and ambiguity[J]. Journal of Zhangjiakou Teachers College, 2003,19(4):24-30.

[8]ZHONG Yixin. Comprehensive information based methodology for NLU: theory and application[J]. INFORMATION, 2004,7(5):641-646.

[9]李 蕾,钟义信. 自动文摘系统中基于全信息词典的复杂语句分析方法及其实现[J]. 电子学报. 2000,28(8):104-106.
LI Lei, ZHONG Yixin. CIL based algorithm simplifying analysis of complex Chinese sentences[J]. Acta Electronica Sinica, 2000,28(8):104-106.

[10]刘建毅,张鹏飞,王 枞. 高性能电子邮件过滤系统的设计与实现[J]. 计算机应用研究,2005,22(4):224-225.
LIU Jianyi, ZHANG Pengfei, WANG Cong. Design and implementation of high performance email filtering system[J]. China Academic Journal Electronic Publishing House, 2005,22(4):224-225.

[11]李 蕾,周延泉,王菁华. 基于全信息的中文信息抽取系统及应用[J]. 北京邮电大学学报,2005,28(6):48-51.
LI Lei, ZHOU YanQuan, ANG Jinghua, et al. Comprehensive information based Chinese information extraction system and application[J]. Journal of Beijing University of Posts and Telecommunications, 2005,28(6):48-51.

作者简介:

李 蕾,女,1974年生,讲师,毕业于北京邮电大学. 主要研究方向为自然语言处理及信息抽取. 发表学术论文多篇. E-mail:lilei@nlu.caai.cn.

周延泉,女,1970年,副教授,毕业于西北工业大学. 研究方向为智能信息处理移动信息服务,发表学术论文多篇.

钟义信,男,1940年生,教授,博士生导师,中国人工智能学会理事长. 主要研究方向为信息科学、人工智能和神经网络,已在国内外发表多篇著作和论文.