中 文 信 息 学 报
第 12 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vo1. 12 No. 1

受限语言子集的理论研究和探索①②

宗成庆 宋 今 陈肇雄 黄河燕

中国科学院计算机语言信息工程研究中心 北京 100083

【摘要】 本文在综述受限语言研究成果的基础上,提出受限语言子集的一种形式化描述模型,并给出其相应的语言特性和数学特性,就受限汉语子集的确定方法问题进行了理论研究和探索。作者希望本文提出的表示模型和确定方法能够引起有关的讨论,并在充分认识受限语言研究的必要性和困难的基础上,将其引向全面深入的发展。

一、引 言

自 1987 年我国有关专家提出受限(规则)语言的概念[1][2]以来,许多专家和学者都对受限语言的研究和发展给予了极大的关注,并进行了不懈的努力,取得了丰硕的成果。鲁([2],1987)对规则汉语(Ruly Chinese)进行了深入有研究,并编纂了《信息处理用规则汉语词汇表》和《信息处理用规则汉语语法基本集》;1991 年南京大学王启祥对受限语言及其设计规则进行了深入研究;陈([3],1991)提出了受限语言子集的概念;1993 年北京大学俞士汶提出了受限的规则汉语概念;1995 年李东提出对汉语受限语言的研究③。所有这些理论和成果都对受限语言的研究,尤其是受限汉语的研究,起到了积极的推动作用。实际上,我国早期在机译研究中提出的"子语言"(Sublanguage)和"局部优化策略"(即"部分优化技术")也是受限语言研究的一种。

由于受限语言在一定程度 上避免了自然语言自身的无限性和复杂性以及当今计算机进行语言信息处理所固有的局限性,因此,受限语言的研究具有毋庸置疑的积极意义。

目前受限语言比较成功的应用主要在自然语言接口的研究和机器翻译研究等方面,包括应用于数据库接口、专家系统接口、CAD/ CAE 系统接口等的语言子集,使用手册范围的机器翻译、天气预报范围用语的机器翻译等。但它们基本上都停留在就事论事,针对具体应用系统的开发,设计具体的自然语言子集上。

在任何一个受限语言处理系统的设计和实现中,语言子集的确定是最基本的工作,子集确定的过小难以达到系统基本的处理要求,而子集确定的过大则又给系统资源造成不必要的浪费,并且影响系统的处理效率。因此,如何将受限语言子集确定在一个合适的范围,

① 本文 1997 年 3 月 11 日收到
② 本文得到国家自然科学基金资助,项目批准号:69772003。
③ 引自俞士汶"关于受限汉语的研究"报告记录,1995。

是实现受限语言处理系统首先应该解决的问题。

本文在继承前人研究成果的基础上,提出了一种受限语言子集的形式化描述模型,并给出其相应的语言特性和数学特性,就受限语言子集的确定方法问题进行了理论研究和探索。作者希望本文提出的表示模型和确定方法能够对受限语言的研究尽微薄之力,并起到抛砖引玉的作用。

二、基本概念和表示

受限语言子集是一个特殊的语言集合,研究这个语言集合是弥补目前计算机理解和处理自然语言"先天不足"的重要途径。本节在引用前人研究成果的基础上提出一种受限语言的形式化描述模型。

2.1 关于受限语言的定义

许多专家都对受限(规则)语言的定义进行过精辟的概括,这里我们分别引用俞([4],1995)和鲁([2],1987)以及国家标准关于受限语言的定义。

定义 1 对某种自然语言的多样性和简约性进行适当的合理的限制,得到的是该自然语言的一个子集,该子集就是受限语言,这样的受限语言又可看作是对自然语言强加一些规则而得到的,所以又叫规则语言[4]。

规则语言(Ruly language)是自然语言的有规则的子集[2]。

国家标准(GB 12200.1 - 90,《汉语信息处理词汇》)曾对受限语言进行过如下定义:

受限语言(Restricted Language)是在词汇、句法、语义及语用等方面受到人为限制的自然语言的真子集。

关于受限语言的定义,可以有两种不同的解释。一种解释认为,受限语言是某种自然语言全集受到人为约束的一个子集。而另一种解释认为,受限语言是由于领域受限而形成的特定子语言。后者在机器翻译、人机接口和自然语言理解等应用领域有着广泛的应用。

从含义上看,笔者认为,受限语言的概念包括两层含义,一层含义是主观对客观的限制,人们为了方便地使用计算机处理某种自然语言而对其人为地设置一些约束条件。另一层含义是客观对主观的限制,即目前现有的技术条件和具体的应用领域对使用计算机信息处理系统的用户有一种约束。上述定义中提到的"规则"不应该仅仅局限于单纯的句法规则或词法规则,而应该包括主客观两方面的一切限制,包括限定的应用领域、词汇量和句型等多种因素的约束,这种约束既遵循了自然语言本身固有的规律,又兼顾了人类使用语言的灵活性和计算机处理语言的可操作性。

2.2 受限语言的形式化描述

为了准确刻划构成受限语言子集各成分的内在联系,便于表达受限语言子集的可计算性,我们提出一种受限语言子集的形式化表示模型。

定义 2 一种受限语言可以形式化地表示为如下五元组:

$$L_R = \{C, V, I, G, R\}$$

其中,C 表示该语言的一个有限的字符集合,C^* 表示 C 中字符构成的所有词汇的集合。如

GB2312 - 80 基本字符集中有 6763 个汉字[5],C * 则是由这 6763 个汉字组成的所有词汇的集合。

V 为 C * 上的一个真子集,用于表示该语言的各种构成成分的名字(例如,单词、短语、句子、主语、谓语等等)。

I 为各种语言成分的语义解释。

G 为该语言所遵循的各种规则集。

R 为受限条件集,通常 R = {限定领域,词汇集,句型集}。

从上述定义可以看出,受限语言子集 L_R 是由约束条件集 R 限定的一个有限的语言集合。确定子集的关键是确定约束条件集 R,即确定限定的应用领域、词汇及其相应的语义和有限的句型。对于受限语言定义的第一种解释,限定的应用领域可认为是无穷大,即不受领域限制。从表示模型可以对受限语言子集解释为:一个有限的词汇集合,这个集合具有以下基本的语言特性和数学特性,使用集合中的任何元素都必须满足 R 中给定的约束条件。

2.3 受限语言子集的语言特性

鲁([2],1987)列举了规则语言的五个基本特点:常用性、有限性、简明性、单义性和规范性。常用性是指词汇、词义、句型等都按其在大量语料统计中的频率来选取,所选的词汇、词义、句型等都是在限定领域内使用频率最高的。有限性是指词汇量和句型量都是有限的。单义性是指选定的每个词的含义是单一的,如果必须保留两个或两个以上的含义时,要作为不同的词区别标记。简明性是指句型表达应该具有简单明了、易于推断的特点,当然,句子结构不应含有歧义性。规范性是指严格遵循句法规则,不允许例外。

实际上,由于受限语言子集是整个自然语言中的一部分,子集中的每一个词汇都可能随着时间的推移和语言的发展而有所变化,因此,受限语言还具有可扩充性和转移性的特点。

可扩充性包括两方面的含义,一方面是指受限语言子集可根据具体限定领域的要求对词汇量、句型量等进行适当地扩充,尤其在科技应用领域,吸收外来词汇和新词是常见的事情。另一方面是指每个词汇的含义和词性可能在应用中得到扩充。例如,在计算机应用领域,电子邮件(e-mail)、互联网(internet)、神经网络(neural net)等都是近几年出现的新词。而"奔腾"一词字典上解释为"(许多马)跳跃着奔跑:一马当先,万马奔腾"(见:《现代汉语词典》,商务印书馆,1992),作动词用,但目前在计算机领域则常用来指一种高档的微型计算机(芯片),作名词用。英文单词"Xerox"本来是一名词,是美国一复印机公司(施乐)的产品商标,可现在该词除了原有的含义以外,又有了动词的含义,常被用来等同于"photocopy",作动词用。

转移性是指某些词汇的词性和含义等随着时间的推移而出现转移的现象,使词汇逐渐淡化了原有的意思而具有了新的含义和词性。例如,"足下"原意是脚底下的意思(特指脚底穿的木屐),而现在则是一种对朋友的敬称(多用于书信)。"集合"一词,字典上有两解释,一个为"许多分散的人或物聚一起:民兵已经在村前集合了。",另一解释为"使集合:集合各种材料,加以分析。"(见:《现代汉语词典》,商务印书馆,1992),两种解释都是动词,而在数学领域内"集合"就是一个纯粹的名词了,特指具有相同性质或某种联系的一组数据或

对象组成的整体。

有关可扩充性和转移性的的例子枚不胜举,在这两种情况下,要使受限语言处理系统准确地运行,并达到处理要求,都必须及时地对系统词典和约束句型作出适当的修改。

2.4 受限语言子集的数学特性

受限语言子集作为一种特殊的数学集合,尽管可以和其它集合一样进行集合运算,但是,许多运算是没有实际意义的,如笛卡尔乘积,闭包运算等。从构造受限语言子集的目的出发,以下给出受限语言子集的一些基本的数学特性和可进行的数学运算。

任意的受限语言子集具有以下数学特性:

·自反性:任意受限语言子集 $A \subseteq A$。

·传递性:若 $A \subseteq B$,$B \subseteq C$,则 $A \subseteq C$。

·反对称性:若 $A \subseteq B$,$B \subseteq A$,则 $A = B$。

设某自然语言 L 的任意两个受限语言子集 A 和 B(R 不同的两个 L_R),其间的运算可以定义为:

·交:$A \cap B$

$A \cap B = \{(f, w, p) \mid A$ 和 B 具有相同的应用领域 f,且任意单词 $w \in A \wedge w \in B$,任意句型 $p \in A \wedge p \in B\}$

·并:$A \cup B$

$A \cup B = \{(f, w, p) \mid$ 应用领域 f 或者为 A 的应用领域,或者为 B 的应用领域,且任意单词 $w \in A \vee w \in B$,任意句型 $p \in A \vee p \in B\}$

·补:$-A$

$-A = \{(f, w, p) \mid f$ 不同于 A 的应用领域,且任意单词 $w \in L$,但 $w \notin A$,任意句型 $p \in L$,但 $p \notin A\}$

·差:$A - B$

$A - B = A \cap \neg B$

三、受限汉语子集的确定策略

根据上述受限语言子集的形式化定义,受限汉语子集的确定包括对应用领域、词汇集和句型集三个因素的确定。应用领域的确定十分简单,而关键是词汇集和句型集的确定。

3.1 词汇集的确定

从自然语言的应用情况看,任何一个受限语言处理系统要处理的词汇集实际上包括两大部分,一部分是日常用语基本词汇,另一部分则是限定领域内的专业词汇。例如,在航空信息服务领域,除了日常用语儿,是“航道”、“航线”等这样一些专业词汇。从集合的构成角度来看,受限语言子集与整个语种的语言集合之间的关系可以简单地表示成图 1 所示的形式。

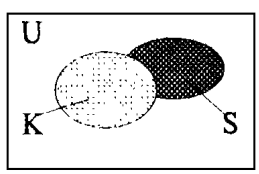

图 1

图中 U 表示某种自然语言的词汇全集,S 表示特定领域的专业词汇子集,K 为日常用语基本词汇子集。那么,整个受限领域的词汇集 $V_R = K \cup S$。

设计词汇子集 K 和 S 的确定算法不仅要考虑选词的准确性,而且必须充分考虑计算机处理的可行性,因为研究受限语言子集的一个主要原因就是要易于计算机处理。

刘(1994)[5]在制定信息处理用现代汉语常用词词表时首次提出了选词函数的术语,并创造性地使用两个选词函数共同选词,从而把选词方法由主观联想为主转化为以定量计算为主。以定量为原则的收词方法客观真实地反映了社会实际用词的规律,避免了传统主观方法建立词典时受专家文化素养、专业学科、社会地位、个人爱好和用词习惯等影响的不足,而且定量计算的方法有利于计算机的自动处理。

北京大学在开发 5 万词条的《现代汉语语法信息词典》时,提出 了电子词典收词的 6 条基本原则和各类词语的具体收词原则[6],台湾中研院也曾开发了 7 万词条的汉语词库,这些成果为受限汉语的词汇选取提供了宝贵的经验。

以下我们针对特定领域中文信息处理系统的开发,提出一个基于词义贴近度的 S 子集词汇确定算法。这个算法的核心思想是:首先由受限领域的专家确定该领域的一个最基本的专业词汇集,称为核心词集,记作 Sc,然后通过求其它词汇与 Sc 中基本词汇的词义贴近度的方法决定是否吸收 Sc 以外的其它词汇。这个算法的前提是已经对所有汉语词汇的词性及特征信息作了标识。

实际上,专业词汇集 S 子集的主要词汇一般是名词和动词,只有少量的数词和量词,而一般不会有形容词、代词和其它虚词,因此,该算法简化为对特定领域专用名词、动词和少数数词、量词的选取。

设所使用的文法特征描述体系对任意单词的标注形式为:

$XP(X_1, X_2, \cdots\cdots X_n)$ …… (1)

其中,XP 为词性,例如:名词(NP)、动词(VP)等;$X_1, X_2, \cdots\cdots X_n$ 依次为对该单词由粗到细、层次逐渐加深的分类特征。例如,对于名词取 n = 5,“飞机”一词的标注可以为:

NP(NCGEN,PY,PYPR,PRTL,TLVI,VIAI)

NCGEN 表示是普通可数名词,PY 表示是一种具体物,PYPR 表示人工生产的东西,PRTL 表示一种工具,TLVI 表示一种交通用具,VIAI 表示用于空中的交通工具。

当然,n 的取值越大,相应的特征信息分类越细,分析的准确性就越高。由于同一个单词可能有两个甚至两个以上的词性,因此,同一单词可能有多个(1)式标记形式。

设 n 个特征 $X_1, X_2, \cdots\cdots X_n$ 的权值系数分别为:$e_1, e_2, \cdots\cdots e_n$,一般地,$1 \le e_1 \le e_2 \le \cdots\cdots \le e_n$。由于受限语言子集内的任何词汇都必须满足单义性的原则,因此,Sc 中的每一个词都只有一种词义标记。任意单词 w 与含 m 个词的核心词汇集 Sc 之间的贴近度 f(w) 通过如下步骤计算出:

Step 1. 从已标注的电子词典中任取一个单词 w(w ∉Sc),如果在 Sc 中找不到与之词性相同的单词,则 f(w) = 0,否则执行 Step 2。直到词典中所有的单词均已计算,转 Step 6。

Step 2. 从核心词汇集 Sc 中任取一个单 w ',设 w 的标注形式为:

$$XP_{W'}(Y_1, Y_2, \cdots\cdots Y_n)$$

w 有 k 个标注,第 j(1 ≤j ≤k) 个标注形式为:

$$XP_{Wj}(X_1, X_2, \cdots\cdots X_n)$$

计算:W 的该标注与 W′的贴近度 $f_j = \sum_{i=1}^{n} e_i \times p(X_i, Y_i)$

其中,如果 $X_i = Y_i$,则 $p(X_i, Y_i) = 1$,否则,$X_i \neq Y_i$,则 $p(X_i, Y_i) = 0$。

以同样方式计算 Sc 中所有单词与 w 的第 j 个标记之间的贴近度。

Step 3. 计算:词 w 的第 j 个标记与核心集 Sc 的贴近度为 $f(w_j) = \max(f_i)$,$1 \leq i \leq m$。

Step 4. 如果 w 还有另外的词义标注形式,则以下一个标注取代当前 w 的标记为新的 j 标注,然后转 Step 2,否则执行下一步。

Step 5. 单词 w 与核心词汇集 Sc 之间的贴近度 $f(w) = \max(f(w_i))$,$1 \leq j \leq k$,转 Step 1。

Step 6. 算法结束。

将所有单词按其与核心词汇集 Sc 的贴近度由大到小的顺序排列,然后根据要求的词汇个数,确定一个贴近度的阈值,所有贴近度大于阈值的单词就是被选中的词汇。

该方法将主观特征标注与计算机定量计算结合起来,只要有一个已经标注的电子词典,就可以实现任何特定领域的词汇子集确定。该方法的困难在于如何较为准确地由粗到细、由浅到深地对各单词进行特征标识,并给出各层次特征的权值系数。

3.2 受限汉语句型的确定

陈([7],1986)给出了确定汉语句型的 3 条重要原则:1. 对立性原则:以语法形式上"有"和"无"之间的互相排斥以及在语法性质上的对立为句子结构分类基础;2. 省略性原则:用简短的形式表示比较丰富的意思;3. 区别性原则:把两种形式相似的句子加以比较,认识它们之间实质上的区别,以此作为归类的依据。

李([8],1986)总结出了 23 类约 429 种现代汉语句型。北京语言学院语言句型小组([9],1989)编写的《现代汉语基本句型》归纳了 219 种句型。唐([10],1993)给出了现代汉语信息处理用简单句的 28 种基本句型。罗([11],1994)提出了以谓语为中心的句型成分分析与句型匹配相结合的分析算法与策略,实现了基于语料库大规模真实语料的汉语句型自动分析和分布统计,为查清现代汉语句型基本状况提供了重要的实践经验和数据依据。他们的工作都为受限汉语句型的研究和确定提供了宝贵的经验。

考虑到受限汉语处理系统中人机互约的特殊性,针对特定应用领域受限汉语处理系统的开发,我们提出如下确定受限汉语处理系统基本句型的基本原则:

1. 以基本句型为主

每一种句型都有它的基本格式,即"规规矩矩"的句型,在实际应用中,基本格式可以转化为其它成型的句子结构格式(转化式),基本格式也可以派生出非成型的结构格式(派生式)[7],句子基本格式是有限的,而转化式和派生式却是无限的,因此,在受限汉语子集的句型确定中,应以基本句型为主要依据,而尽量避免多义句和同义句。例如,句子:"下个周我有可能去纽约",该句是两个 VP 共一主语的"连动式",多用于正式口语,是基本格式,它可以转化和派生出若干意思相同或接近的句子,如:"下个周我有去纽约的可能","下个周我去纽约是有可能的"等等。基本格式应是确定句型的主要依据。

2. 以高频句型为主

根据受限领域内大量语料统计的结果,以使用频率较高的句型为主,对于一些使用频

率较低和极少使用的句型,尽量不予考虑。

3. 以简单句为主

无论是数据库查询接口、机器翻译等以文本形式进行处理的自然语言处理系统,还是语音翻译、人机接口等以语音形式进行处理的自然语言处理系统,确定受限汉语句型的根本目的是为了便于计算机理解,而复杂的句子必然会给系统分析机制带来许多困难,因此,确定受限汉语句型应以简单句为主。

4. 以句子表层含义和常用意思为主

汉语句子的含义往往是丰富多彩的,即使简单句也不例外,在确定受限汉语句型时应以受限领域内句子的表层含义和常用意思为主,而一般不考虑句子的深层含义和言外之意。当然,句子结构更不应该含有歧义。

5. 以规范句型为主,但不排斥非规范句型

确定汉语句型应以符合汉语语法的规范句型为主,但考虑到具体的受限领域和计算机处理的方便性,并不排斥在必要的时候引入个别的非规范句型。实际上,好多语法本身就是人们根据语言的特点而人为制定的,在实际交流中,尤其是日常口语交流中,也并不是人们讲出来的每一句话都符合句法规定,但丝毫不影响信息的交流,因此,在确定受限领域句型子集时,不应该排斥非规范句型的存在。

四、问题讨论与结语

受限语言的研究是一项复杂的工作,其中既有语言学方面的问题,也有数学方法和计算机处理技术等方面的问题,许多理论和技术问题还有待于进一步研究和探索。我们认为,以下两个问题不容忽视:

·自然语言的模糊性处理。自然语言本身具有难以完全量化的模糊性,无论是日常用语词汇和受限领域专业词汇的确定,还是基本句型集的确定,都具有一定的不确定性,如果能够利用模糊数学方法,例如通过构造隶属函数的方法确定词汇子集,在句型确定中引入模糊推理和不确定性推理机制等,将为受限语言子集的研究提供一条重要的途径,但是如何把模糊数学方法与语言学知识结合起来有待于进一步探讨。

·把人对语言的限制与语言对人的限制结合起来,语言是人类在社会实践中自然形成的,然而人类在使用语言的同时又人为地设置了某些限制,比如,语法规则,书写格式等,这些规定在很多方面有助于计算机进行语言处理,但有时也阻碍了计算机对语言的处理,因此,在尊重人们语言习惯的同时,在必要的时候可以考虑对语言使用者进行某些约束,以有利于计算机对语言的处理。

我国许多中文信息处理专家都对受限语言的研究提出了很好的指导性建议。随着计算机技术的发展和人工智能等相关理论研究的不断深入,我们相信受限语言的研究必将取得更大进展。本文中提出的受限语言的数学模型和词汇集的确定方法以及确定句型集的基本原则是作者在受限语言的语音翻译系统研究中得到的一些基本认识,我们希望本文提出的问题能够引起相应的关注,并将受限语言的研究推向深入。

致谢 作者衷心地感谢本文审阅者提出的宝贵建议和修改意见。

参 考 文 献

1. 陈力为,当前中文信息处理中的几个问题,计算机世界,1987 年第 21 期第 34 版。
2. 鲁川,信息处理用规则语言,中文信息学报,1987 年第 1 卷第 4 期。
3. 陈肇雄,梁南元,自然语言处理理论研究发展战略,《计算机科学技术》,自然科学学科发展战略调研报告,科学出版社,1994。
4. 俞士汶,关于受限的规则语言的设想,语文现代化论丛,山东教育出版社,1995。
5. 刘源,谭强,沈旭昆,信息处理用现代汉语分词规范及自动分词方法,清华大学出版社,广西科学技术出版社,1994。
6. 王惠,朱学锋,俞士汶,《现代汉语语法信息词典》的收词原则,中国计算机报,1994 年第 21 期,79 - 83 版。
7. 陈建民,现代汉语句型论,语文出版社,1986。
8. 李临定,现代汉语句型,商务印书馆,1986。
9. 北京语言学院句型研究小组,现代汉语基本句型,世界汉语教学,1989。
10. 唐泓英,姚天顺,王宝库,关于汉语句型,中文信息学报,Vo1. 7,No. 1,1993。
11. 罗振声,郑碧霞,汉语句型自动分析和分布统计算法与策略的研究,中文信息学报,Vo1. 8,No. 2,1994。

The Theoretical Research and Probe into Restricted Language

Zong Chengquing, Song Jin, Chen Zhaoxiong and Huang Heyan

Research Center of Computer & Language Information Engineering.

The Chinese Academy of Sciences, Beijing 100083

Abstract After summarizing some important research results on restricted language, this paper proposes a fomalization model for the restricted language, and also discusses the linguistic characteristics and mathematical characteristics of the restricted language. The paper gives a method to determine the vocabulary and sentence patterns of a restricted language subset. The authors hope the model and the method presented in this paper evoke a heated discussion and propel the approach to the restricted language.

Keywords: Restricted language, Sublanguage, Ruly Chinese