

小学语文 ICAI 系统诊断模块中的 造句错误类型分析^{*}

杨开城

北京师范大学现代教育技术研究所 北京 100875

摘要 本文从句法分析的角度出发,全面分析了错误文本的技术表现,从信息处理的需要出发扩展了造词法的概念,提出把词语搭配错归入句法错的观点,并试着对“语义错”这种模糊的提法给出了明确的描述。本文在详细分析各种错误类型的特征的基础上,提出了相应的诊断策略。

关键词 造词法 搭配 语义

一、错误类型分析意义及分析的角度

错误类型分析对于小学语文 ICAI 系统造句诊断模块的目标设计以及功能评价都是一个至关重要的参照。如果分析得不清晰不彻底,就会妨碍造句诊断的总体功能设计,甚至影响整个 ICAI 系统的性能指标,而且会使 ICAI 系统智能诊断功能的评价指标变得模糊而缺乏可操作性。

对错误类型的分析通常有两种角度:人的角度和计算机算法分析的角度。从人的角度对错误进行分类,虽然有一定的客观性,但会出现同一类错误在系统中会对应多种技术表现的现象,并且处理方式不同的系统技术表现也会迥异。这就给系统功能设计和最终的性能评价带来很多不便。从计算机算法分析的角度进行错误分类,会因语言处理系统的用途不同而出现差异。但从计算机处理的角度进行错误分类的一个明显优点是利于界定和评价系统的功能。

我们正在开发的小学语文 ICAI 系统中,造句错误诊断模块与常规的文稿校对技术在很多方面(处理文本的性质、诊断错误类型的重点等)存在着差异。我们认为有必要根据系统的特殊要求对错误类型从计算机处理的角度进行系统地分析。下面将讨论分析结果。

二、错误类型分析

我们粗略地将错误类型分为不成词错、造词法错、句法错、语义错和百科知识错五大类。

2.1 不成词错

由于汉语在书写时汉字是连续排放的,词与词之间没有分隔标志,所以在对汉语句子进行分析之前,必须对句子进行分词处理,将汉字串切分为词串。汉语词汇中有一些是多音节单语

^{*} 本文于 1998 年 10 月 27 日收到

素的单纯词,并且构成语素的字在任何情况下都不能独立成词。如果切分过程中出现不成词的“语素字”,就应判定句子出现了不成词错。如:

吐鲁番的葡萄熟了。(本来是缺了一个“萄”字,但技术表现则是切分出来一个不成词的字)

除了有些多音节单纯词在丢字、多字或错字的情况会出现不成词错外,有些带有粘着语素的派生词和复合词也会造成不成词错。如:

他碌了一整天。(少了一个“忙”字,“碌”为粘着语素,不能独立成词)

由于各个系统的收词量不同,因此系统应根据各自的切分算法和收词情况,在采取切分歧义消解措施的情况下,如果切分出来的字无法在词库中检索到,就应诊断为不成词错。如果不采取消歧措施,下面的句子会被认为是错的(假设按最大匹配算法分词):

做了解释(“释”在现代汉语中极少单用,因此系统会诊断该句为不成词错)

考虑到小学生的用词量,我们的系统中单字词只收入 1378 个词条,因此,6763 个汉字中就有 5385 个汉字不能用来充当单字词,我们认为这些汉字从小学语文教学角度来说都属于语素字(尽管从学术的角度来说并不严格)。

2.2 造词法错

这里所讨论的造词法概念与传统意义上的构词法有很大的差异。传统的构词法指的是汉语词汇的构造规律,包括单纯词、派生词以及复合词。本文所讨论的造词法尤指汉语词汇在原型的基础上按某种规律进行变化而形成新词的现象,它包括:加缀法、重叠法以及内部扩展法。

加缀法是指在原词语的前或后或中间的某个位置加上词缀而构成新的词语。如:第十、看得见、物理学家、同学们等等。由于这样使用的词缀都可单独切分,所以无法在分词阶段诊断加缀法的错误。由于汉语中这样的词缀为数不多,并且大多数词缀的结合能力比较低,因此本系统对加缀法构成的词一部分直接收入词库,另一部分归入句法规则(类似于 PFF + NNCO、NNCO|NNCQ|NNAB + SUF 的形式,其中 PFF 表示前缀字,SUF 表示后缀字),不在分词阶段处理。所以造词法错不包括加缀法错的情况。

重叠法是指某些词(尤以单音形容词、单音动词和双音动词居多)在原型基础上按某种规律重叠而形成新词的现象。如:“看看、红红的、研究研究、试了试”等。重叠方式依词类和词本身而定,包括 AABB 式、ABB 式、AAB 式、AA 式、A 了 A 式、A - A 式、A 了一 A 式等等。新构成的词语在严格意义上有时是词组,但其模式是固定的,且构造极有规律,日常使用时也是作为一个整体来使用的,因此为了处理方便,我们仍视它们为词。

我们的系统对动词和形容词重叠的处理是不同的,由于形容词重叠后可能会发生词类变化,有的变成了状态词,有的则变成了副词,并且那些可重叠的形容词数量不大(大约只有 80 几个),所以我们将这些重叠形容词直接收入词库。我们将动词的重叠方式分为三类:自身重叠式,包括 A[了][一]A 式、ABAB 式;AAB 重叠式和 AABB 式。对于第一类重叠我们将其归入句法规则,只需一条规则即可。后两种重叠在分词阶段处理。

为保证分词和句法分析阶段中能正确处理重叠现象,词库中标有词条的重叠模式信息,如果分词或句法分析阶段发现词的重叠现象不符合词库中所标记的模式,则认为出现造词法错。

有些学者认为加缀法和重叠法不属造词现象,而是一种词的形态变化。笔者认为,所谓“形态变化”这一术语来自印欧语语法理论体系,它是有特定涵义的。词的形态变化是为了表达某种语法功能,如人称、性、数、格、时、态和体等。其目的是为了使句子中各成分间的语法功能达到某种一致。词的形态变化并不会带来词概念义的变化,更不能表现某种风格意义的变化。汉语词形(与词音、词义相对而言)的变化,尤其是加缀法,虽然非常类似于印欧语中的形

态变化,但本质上来说是两种截然不同的语法现象。汉语词形的变化通常会带来词概念义、风格义,甚至会引起语言单位的变化(由词变为词组)。并且汉语中人称、性、数等观念虽然也存在,但主要依靠词义和虚词来表现,单从词形变化上看不出其上述语法功能的变化。因此不考虑功能和意义上的差别,只从形式上的相似就类推汉语中存在形态变化的论断是片面的。

但是,我们应看到,词形的变化,尤其是词的重叠和词的内部扩展,都会带来词义的变化和在句子中组合功能的变化。实际上成了一种新词(或分词单位)。所以我们将这种词形变化归入造词法之列。

词的内部扩展是指某些双音节离合动词中的一个字或两个字都不能单独成词,但却可以在中间插入一些特殊成分而构成一种词组来使用。如:鞠了一躬、照了一张相。由于“鞠”、“躬”、“相”在现代汉语中极少单用,因此如果将其切分开,必然会出现不成词的情况。所以这种现象最好不要在句法规则中处理。我们看到,这种可扩展的离合词其扩展的模式是有限的,而且很有规律。所以我们将这种现象归入造词法之列,在分词阶段处理。词库中将可扩展的词(主要是VG\$1类动词)标明扩展的模式,在分词时检查词串是否是扩展模式(如A了一B),再检查AB这个词是否可以扩展以及扩展的模式,如果这个词不可扩展或者不支持当前的扩展模式,则认为出现造词法错。

目前我们只处理以下六种扩展:加‘了’、加‘过’、加‘着’、加‘个’、加‘一’和加‘的’。

由于重叠法和扩展法都有模式可遵循,因此,使用模式匹配的算法就可诊断出重叠法和扩展法的造词错。我们所要做的只是在词库中相应的词条上标注出重叠规则和扩展规则即可。下面列举几例造词法错:

我的的书包不见了。(的不可重叠)

他鞠着躬。(不符合“鞠躬”的扩展模式)

教室被打扫得王王净。(AABB式重叠模式中少了‘B’)

综上所述,根据不成词错和造词法错的特点,两者大多情况下都是在分词阶段处理的,不能等待句法分析器来诊断,而保证正确处理造词法现象的基础是词库中的造词法模式信息。

2.3 句法错

为了讨论方便,这里的句法规则也包括了词法规则。由于汉语的词组构造与句子构造在很多情况下是非常一致的,所以有时单从结构上无法判断一个语言片断是词组还是句子。从汉语语法学的研究现状来看,尚无一个统一的标准来对词组和句子进行划界。甚至连对句子的定义也存在争议。而目前计算机语言处理技术,由于没有语义理解的帮助,多数以标点符号间的语言片断而不是以句子为分析处理的最大单元。因此,计算机中文信息处理系统中多数不区别句法规则和词法规则,而是将两者合起来统一处理。但笔者并不同意“词组本位”的观点。尽管汉语的词组与句子在某种情况下是同构的,但毕竟两者位于不同的语法分析平面,而且汉语中有很多句式无论如何也不能看作是词组。笔者认为,句子和词组的一个最大区别并不在于结构本身,而在于结构关系的紧密程度和灵活程度。词组中支配单元和被支配单元的位置关系是比较固定的,而句子中这种支配和被支配单元之间的位置关系就比较灵活了,如:宾语和定语前置现象等等。80年代兴起的配价语法的研究成果也说明了这一点。虽然我们在讨论时对词法规则和句法规则统称句法规则,但我们系统内部对词法规则和句法规则的处理是有差别的。这主要表现在对搭配检查类型的动态生成上。

任何语言的句法系统都包括两大部分:句法的结构系统和结构关系系统。前者描述符合语法的语言片断的结构以及结构成立的句法限制条件(限制条件越丰富,句法分析的精度就最

高) ;后者描述结构中各单元间的语义支配关系,在句子层面上结构与结构关系并不是一一对应的,而是一对多的关系。因此,句法错表现为两个方面:

1. 句法结构不合理

这种不合理或表现为句法结构有缺欠或表现为违反句法约束条件。如:

我们狠狠地打击了。(“打击”是必带宾动词,因此,该句缺少宾语。可以续接“敌人”一词)

这首歌很耳熟。(“耳熟”的配价为2,要求有两个必有的名词性成分同现。可在“很”前加“我”)

他正在睡觉着。(“睡觉”已有“进行、持续”的含义,因此不能加“着”,应删掉)

我商量了一下。(“商量”的配价为3,要求前面的主语表复数,可改“我”为“我们”)

句法结构不合理很容易被发现,所有的句法规则匹配失败就表现句法结构不合理。更困难的是句法约束条件的检查。因此,句法分析的核心技术应是句法约束体系的设计和实现。由于汉语词汇缺乏形态变化,无法直接从词形提取句法结构的约束信息,只能在词库中进行结束信息的标注。我们目前只有400多条句法规则,尚未覆盖全部语言现象,但足够小学语文教学使用了。在这400多条句法规则基础上,我们提出了一套用于限制句法规则生成能力的句法语义特征体系。之所以将其称之为句法语义特征体系,是因为这些特征来源于词汇的语义,但却直接影响着句法规则的生成能力,介于句法和语义之间的层次。我们的句法语义特征体系将句法语义特征分为11类,包括动词、助动词、形容词、副词、名词、代词、时间词、量词、介词、连词和处所词。每种类型的句法语义特征的细分类各不相同,最多的动词有84项,最少的连词只有3项。需要指出的是,我们的句法语义特征体系不是对词汇进行义素分解的结果,虽然体词的句法语义特征借鉴了义素分解的方法,但提出整个句法语义特征体系的目的是构成对句法规则的约束而不是句子的语义解释(尽管义素分析法无法达到句子语义解释的目标)。

我们的句法语义特征共分三个层次。第一个层次直接标注于规则中,代表词汇单元(不是规则单元)对句法规则的选择性,如在规则 $V_{GNC} < LeAttach > 0 > +$ 了 + $NNCO$ 中(经过简化了), $LeAttach$ 项是指动词是否可以后接助词“了”。在这条规则中 V_{GNC} 单元对应的词汇的句法语义特征中的 $LeAttach$ 项必须大于0。第二层次的句法语义特征也标注在规则中,这规则单元之间语义支配关系明确的情况下使用。比如动宾词组中的动词和宾语之间要有第二层次的句法语义特征检查,检验宾语是否满足动词对宾语的要求。第三层次的句法语义特征并不在规则中进行标注,也不在句法分析过程中进行检查,而是用于在句法分析结束后进行句法成分结构完整性的检查,比如是否缺少宾语或主语等等。

词库中的每个词条都有三种句法语义特征的标注,一种是词汇自身的句法语义特征,另一种是该词条在应用于句法规则时的条件约束信息,第三种是该词条可能提供的句法语义特征。所以,我们词库中句法语义特征标注不是静态的,而是动态的,只有这样才能发挥句法语义特征体系的约束作用。

下面列出了我们句法规则的两条例子。

$DB_VP = (\{ V_{GNC} | V_{GNM} \}) < \$V_G \$, DeCompAttach > 0 > @c @d1 + NNCO | NNUC | NP @p1 + V_{GNC} | V_{GNM} < R > + AUCM + AJM1 |$
 $AJM2 < \$AJ \$, DeModify > 0 >$
 $ZDB_VP = PRPD | PRPQ < \$PR \$, DiVAttach < 2 > + (DB_VP | ZDB_VP) @c$

2. 词语搭配错

即句法结构关系不合理;如:

温暖如春的阳光。(“温暖如春”修饰“阳光”不恰当,应删去“如春”)

我们喜爱祖国。(“喜爱”的宾语不能是“祖国”,应改为“热爱”)

对于搭配是属于句法范畴还是属于语义范畴的争论,自乔姆斯基的转换生成语法问世就开始了。乔姆斯基一开始认为搭配纯属词义问题而与句法无关;后来又改变了他的观点,把搭配问题放在句法中处理。方法是每个词提供它的句法特征,词与词组合成词组时要进行句法特征一致性的检查以判断是否符合语法。这种做法类似于格语法的格框架。但单从词的句法特征出发并不能解决全部搭配问题。这主要是因为词的句法特征是基于单个词的,不可能描述得过细。此外,词的句法特征在处理词组与词组之间的搭配时显得无能为力。因为词的某些特征是不可以被词组继承的,并且词组又会生成多种多样的句法特征,有时涉及的是非语言的知识。麦考莱曾举了一个例子来反对乔姆斯基的观点:That electron is green. 谁也不会为 electron 增加它是什么颜色之类的句法特征。麦考莱又指出,“bachelor”和“unmarried man”在搭配上是一致的,但其句法特征却是不同的。因此,他认为搭配实质上是一种语义选择限制,而不是句法选择限制,搭配应属于语义范畴。

历史上的这场争论最终未取得共识,其根本原因在于对什么是语义、句法中应有多少语义成分参与等理论观点有分歧。笔者认为,以上争论似乎不是概念本身的争论,而是一种搭配问题能不能用句法手段来处理的争论。似乎如果用句法手段能解决搭配问题,搭配就应属于句法范畴,否则就应属于语义范畴。难怪乔姆斯基最后认为,某一语义现象能不能进入语法并没有先验标准,只有经验标准。

笔者认为,在把句法和语义结合起来研究已成为一种必要趋势的今天,讨论这个问题的意义并不太大。搭配问题本质上属于语义现象。但这并不说明搭配问题就得在语义系统中处理。我们不能根据某一语言现象的根源是语义就判断该现象属于语义。对于计算机语言信息处理来说,必须决定搭配问题是作为句法现象在句法结构中检查,还是作为语义现象在语义结构中检查。正如中文信息处理用的词库收词一样,有些“词”严格来说并不是词,却收入词库,其目的是为了便于处理。对于搭配问题也是一样,需要人为地加以选择。我们认为绝大部分搭配现象都可以在句法层次上处理,所以我们选择前者,原因有三:

(1) 由于汉语无形态变化,主要靠词序来表现句法单位之间句法关系。因此汉语的句法结构普遍存在着结构层次歧义的消解问题。如:“ $P + NP_1 + \text{的} + NP_2$ ”的结构层次是“ $P + (NP_1 + \text{的}) + NP_2$ ”还是“ $P + (NP_1 + \text{的} + NP_2)$ ”的问题就是一种典型的句法层次歧义的消解问题。对于这类问题的解决办法大致有二:一是对词类进行细分,但单靠词类细分并不能解决全部问题;二是利用搭配检查。也就是说,一种句法结构是否合理不仅仅取决于结构本身,还取决于结构层次关系的合理性。从这个角度看,搭配检查是位于句法层次的操作,应作为一种句法手段的形式存在。

(2) 有些搭配现象完全是约定俗成的。在语义上无法找到根据。

(3) 有些搭配现象表面上是语义现象,但实质上与百科知识有关,因此,在不理解百科知识的情况下是无法进行检查的。从信息处理的角度,我们把这类现象排除在搭配之外。

我们在句法分析过程中所检查的搭配并不局限于词组内部,而是包括了句子成分间所有的语义支配关系的检查。林杏光先生曾正确地指出,搭配是一个系统。从信息处理的角度看,搭配检查包括两层意思:能否搭配以及搭配成立的限制条件。搭配的限制条件是非常复杂的,有的涉及词义,多数涉及句法结构。因此类似于格框架的静态标注无法对其进行全面的描述。因为搭配限制一部分来自于词义本身,一部分又受句法结构的制约。所以,我们除了在词库中加入能独立起作用的搭配限制条件外,大部分限制条件都标注在词语搭配词典中了。这些限制条件大致分为以下几类:词语适用的句式结构限制(如有些动词在某些情况下不能用于“把”

字句)、搭配单元之间的位置关系限制(如有些相向动词要求它的一个施事以介词引进并放到前面)、搭配单元自身结构复杂性限制(有些词要求它支配的中心词必须带有其它附带成分而造成复杂结构)、搭配单元在进行搭配时中心词的暂时转移限制(有些动词在支配包含数量词的名词短语时,有语义支配关系的并不是中心名词,而是修饰该名词的容量量词)等等。

从我们对搭配限制条件的分类来看,80年代兴起的配价语法的研究成果将非常有助于对总结搭配限制条件的总结。此外,汉语词汇学研究,尤其是动词、形容词、副词以及各种虚词的语法意义的深入研究将会使我们系统中的搭配限制条件越来越丰富和精确。

构建一个词语搭配词典是件浩大而复杂工程,即使基于封闭语料的搭配词典的建构也是相当繁重的。虽然我们在这方面进行着尝试,但目前我们尚未建成能够应用于实验的搭配词典。根据我们的经验,前文所提及的句法语义特征体系基本上可以作为搭配词典中的搭配条件来使用。

2.4 语义错

有很多有关文稿校对的论文在分析文本错误类型时,都将语义错单列。但大都只列举一些实例而非从句子语义的内涵或表征的角度来说明什么是语义错。按本文的观点,那些实例大多属于词语搭配错。关于什么是语义,哲学家、逻辑学家以及语言学家们都从不同的角度提出了不同的看法。这场发生在不同领域之间的辩论一直持续到现在。笔者认为,与其说难以对语义的内涵加以描述,倒不如说尚未找到一种合适的方法来表征语义。也就是说,任何一家理论在试图表征(包括形式化和非形式化的)语义时都遇到了语义中无法表征的部分。但大多数学者都认为,语言的语义并不等同于语言所表现的百科知识。句子的语义并不是句子所传递信息的全部,尽管语义充任着非常核心的角色。阅读心理学研究表明,人们在阅读一个句子时,进入长时记忆系统的并不是句子的具体字和词,而是句子所蕴含的语义信息,并且语义是以命题以及命题网络的方式存贮的。因此,我们认为中文信息处理领域中句子语义的表征也应应以命题为基础。当然命题并不是句子语义的全部。句子的意义包括语法意义和语汇意义,命题只能用来表征语汇意义。此外命题表征还存在着如何表征命题谓词和主目意义的难题(有学者试用义素分析以及复杂特征集合—运算的办法来弥补命题表征的不足)。尽管如此,我们认为句子的命题以及命题网络仍然是诊断语义错的基础。但我们并不能从一个句子命题真伪来推断句子语义是否有错。一个句子命题可以真,也可以假,还可以无所谓真假。比如:“分子由原子组成。”和“分子不是由原子组成的。”这两个句子一个是真命题,一个是假命题,但都无语义错。

笔者认为,一个句子的语义错并不表现为命题的真伪,而是表现为句子所蕴含命题间的矛盾性和不协调性。自相矛盾的句子是语义错的典型例子。如:“一个喧嚣而寂静的夜晚。”这个句子包含有两个基本命题:喧嚣(夜晚)和寂静(夜晚)。但这两个命题是矛盾的、不相容的。所以这个句子有语义错,而无搭配错。由于搭配在命题中有时表现为谓词与主目之间的关系,但这种关系的检查已在句法分析阶段完成了。因此,这里的语义只表现为命题之间的矛盾性和不协调性。

获得句子的命题表征并非易事。这项技术也是计算机语义处理的一个难题。以 Montague 语法为代表的逻辑语法是目前获得句子的语义解释的最有潜力的语法理论,这些语法理论走向实用还需要时间。我们的 ICAI 系统同样不能处理这类错误。

2.5 百科知识错

百科知识错包括事实错、知识错(概念、原理性错误)、修辞错等等。这类错误,尤其是修辞

错,与语义错似乎很难区分或者说有交叉。但这类错误所涉及的不是语言知识运用问题,而是非语言知识的运用问题。对于百科知识错,面向一般应用的系统是无能为力的。因为它涉及的非语言知识面太广,系统无法容纳。

综上所述,在我们的小学语文 ICAI 造句错误诊断模块中,句子的错误类型被划分为不成词错、造词法错、句法错(包括句法结构错和词语搭配错)、语义错以及百科知识错。这里并未包含标点符号错。这是因为标点符号错在技术表现上不是具有语义上的歧义,系统无法诊断;就是相当简单,没有学术研究价值。并且标点符号并不是小学生造句时最重要、最常见的错误。目前我们的系统还不能处理语义错和百科知识错。由于我们词语搭配检查归入句法分析器中,因此,我们的句法分析是完整的,可望在句法分析阶段就能诊断出绝大部分造句错误。

参 考 文 献

- [1] 侯敏. 'P + NPI + 的 + NP2' 结构的分化处理. 语言文字应用, 1998 年第 1 期
- [2] 陈小荷. 一个面向工程的语义分析体系. 语言文字应用, 1998 年第 2 期
- [3] 赵新. 动词重叠在使用中的制约因素. 语言教学与研究, 1994 年第 3 期
- [4] 林杏光. 论词语搭配及其研究. 语言教学与研究, 1994 年第 4 期
- [5] 杨成凯. 关于短语和句子的构造原则的反思. 汉语学习, 1993 年第 2 期
- [6] 周国光. 现代汉语形容词配价研究述评. 汉语学习, 1995 年 2 期
- [7] 周国光. 现代汉语动词的配价研究. 汉语学习, 1996 年第 1 期
- [8] 慕勇, 孙才, 罗振声. 汉语文本自动查错与确认纠错系统的研究. 计算语言学进展与应用. 北京: 清华大学出版社, 1995
- [9] 石安石. 语义研究. 北京: 语文出版社, 1994
- [10] 徐烈炯. 语义学. 北京: 语文出版社, 1990
- [11] 刘叔新. 汉语描写词汇学. 北京: 商务印书馆, 1990
- [12] 范晓, 杜高印, 陈光磊. 汉语动词概述. 上海: 上海教育出版社, 1987

Error Type Analysis of Sentence Error Diagnosis Module in ICAI System for Chinese Language Learning in Primary School

Yang Kaicheng

Institute of Modern Educational Technology Beijing Normal University 100875

Email: yangkc@elec.bnu.edu.cn

Abstract In this paper the author analyzed those technical expression of erroneous sentences from the point of view of syntax analysis and extended the concept of word-building rules to meet the need of information processing. The author, trying to describe semantic errors clearly and detailedly, believes that collocation errors should belong to syntax errors.

Key words word-building rules collocation semantics