词汇语法拼写校对软件——功能语法的应用实例

许庆欣

(厦门大学 外文学院, 福建厦门 361001)

摘 要: 真词错误拼写校对是自然语言处理领域中的一个难题。真词错误具有出现频率高、种类多样的特点。现有的计算机拼写校对软件不足以侦别所有的真词错误。词汇语法拼写校对软件模型以系统功能语言学理论为指导, 有四个处理单元, 分别对应语言的四个级阶, 可以侦别出词汇层、局域性句法层、整体性句法层和语义层次上的拼写错误。词汇语法拼写校对模型在进行真词错误纠错工作的时候充分考虑了文本的语境知识, 最终产生按照盖然率大小排列的校正参考。

关键词: 真词错误; 词汇语法; 拼写校对

Abstract: Real-word spell checking is a long-existing problem in the natural language processing field. This paper argues that real-word spelling errors are much more common than expected and the existing computer spelling checkers are not efficient enough to tackle all real-word errors. The lexicogrammar computer spelling checker model is based on systemic-functional linguistic theory. With four processing units which correspond to the rank scales in language, the proposed model can detect spelling errors at the lexical level, local and global syntactic level as well as semantic level. The lexicogrammar-based model takes textual context knowledge into consideration, and finally produces a probabilistic list for correction candidates.

Key words: real-word error; lexicogrammar; spell checking

中图分类号: H08 文献标识码: A 文章编号: 1008-665x(2007)02-0049-06

1 引言

自从书面语言诞生以来, 拼写校对就一直伴随着人类。随着文明的进步和技术的现代化, 拼写校对成为一个日益普遍和现实的问题, 不仅困扰着人类, 也困扰着计算机。真词错误的出现频率远远高于人们的想象, 而现有的计算机拼写校对软件不足以侦别所有的真词错误。本文借鉴系统功能语法的相关理论对现有的计算机拼写校对软件进行改进, 提出一个新的计算机拼写校对软件模型——词汇语法拼写校对软件模型, 以输入文本中的真词错误为纠错目标。

词汇语法拼写校对软件模型是以词汇语法为基础, 针对真词错误的一个理想的计算机拼写校对软件模型。该模型中有四个处理单元, 可以侦别出词汇层、局域性句法层、整体性句法层和语义层上的拼写错误。词汇语法拼写校对模型在运行过程中充分考虑了文本的语境知识, 因此在词汇语法拼写校对模型基础上编写的计算机拼写校对软件理论上可以解决真词拼写错误的侦别问题。

2 词汇语法拼写校对软件的意义

提出一个新的计算机拼写校对软件的原因有两个: 一是现实文本中真词错误出现频率高、

收稿日期: 2006-10-12

作者简介: 许庆欣(1978-), 女, 助教, 研究方向: 功能语言学

类型多样,难以把握;二是现有的计算机拼写校对软件不具备足够的语言学知识,在真词错误面前力不从心。

2.1 真词错误

真词错误是指错误的词汇是词典中的词,但它与上下文搭配不当,如 Their she goes. 中的 Their 应为 There。Baddeley(1980:456－462)通过研究发现,在手写文本拼写错误中约 30%是真词错误。Eastman 和 McLean(1981)及 Young(1991)的研究表明,在计算机自然语言语料库中真词错误至少占 25%。Peterson(1986:635)的结论是,真词错误普遍存在于短的、常用的词汇中,如 sat 就经常被错写成其他真词(set, sit, sad 等)。因此,理论上在键盘输入文本中真词错误的频率高达 16%。从自然语言处理角度看,真词错误可以分为以下几个层面:局域性句法层(local syntactic level)、整体性句法层(global syntactic level)、语义层(semantic level)、话语层和语用层(discourse level and pragmatic level)。

产生真词错误的原因多种多样,包括语音影响(如 piece/peace)、相似词影响(如 His name was Mrs. Williams.)、拼写干扰错误(orthographic intrusion errors)等。键盘的布局也是造成真词错误的重要原因。研究(Gnudin, 1983)表明,在所有的字母替换错误中 58%与键盘邻近键有关,如 I bequeath him this property to have and go hold.(go 应为 to)另外,还有一些是常见的拼写错误,如 compliment 与 complement 经常被错误地互换。

2.2 现有计算机拼写校对软件的评估

真词错误的校对必须依赖上下文,所以它又称为上下文相关的词校正(context－dependent word correction)。目前英文的拼写检查系统都根据各种错误来源将彼此容易混淆的词收集到一起,形成混淆集(confusion set)。一个混淆集合 S={W[,1],W[,2]…,W[,n]}表示 S 中的每一个词 W[,i]都容易在使用中与 S 中其他的词发生混淆。这就意味着当我们在待校对文本中遇到一个 S 中的词 W[,i]时,就要考虑 S 中其他的词是否更适合该处的上下文。例如,当遇到 desert 或 dessert 时,就要考虑究竟是 dessert 还是 desert 更适合该处的上下文。此时,英文真词错误的校对模型就转换成一个排歧的模型。模型的核心任务就是根据上下文从混淆集中选择一个最合适的词(张磊,2000)。

上下文词和同现的方法分别利用了词之间有序和无序的依赖关系。为了综合两种方法的优势,Yarowsky(1994)和 Golding(1996)分别提出了几种不同的混合模型。这些模型的出发点是将上下文词和搭配统一看成特征。所谓特征,就是在特定目标的上下文中出现的语言现象,无论上下文词还是搭配,都是特征的一种。关于英文校对更多的细节详见 Kukich(1992)。

通过对现有的几种主要计算机真词错误拼写校对软件进行评估,发现它们在计算机语言基于规则的方法和基于统计的方法方面有很大突破,但由于缺乏语言学理论的支持,真词错误的侦别能力相当有限(见表一)。

3 词汇语法拼写校对软件的构成

从语言学理论上看,本文提出的词汇语法拼写校对软件是以系统功能语法的词汇语法、系统网络、语境和盖然率为理论基础。语言是意义潜势(meaning potential),是系统资源(systematic resource)。因此,语篇分析的基础是对语言系统的研究,对于语言的描述可以看成是对选择的描述。系统功能语言学家就是分析既定语境下语言使用者的多种选择,进而得出具体的语言产品。选择是由语境决定的,而且可以放在语言的多个层次上进行分析。

从计算机科学方面看,该校对软件结合了计算机科学中基于规则的方法和基于统计的方法。我们的假设是每个英文单词都可以被误认为是其他真词。因此,在进行拼写校对的时候应该在各个语言层次上进行检查。同时,人脑在自然语言处理上采用的是平行处理方式,计算机自然语言处理软件也应如此。也就是说,在词汇语法的各个层次中,计算机拼写校对软件要从高等级检测到低等级,反之亦然。

词汇语法拼写校对软件的输入是词汇流,因此该软件不是在独立词汇层次上工作,而是把校对对象放在语境中,在词组、小句、句子甚至语篇层次上进行侦别。因此,理论上该校对软件可以成功地识别出整体性句法层、语义层或话语层上的真词错误。该软件的输出是按照

表 1　现有计算机拼写校对软件真词错误的侦别能力

Factors		词汇层错误	句法错误		语义层或语用层
			局域性	整体性	
主要计算机真词错误拼写校对软件	句法方法 (CRITIQUE) ①	是	是/否	否	否
	词性法 (part－of－speech method) ②	是	是/否	否	否
	词汇组合法 (word combination method) ③	是	是	否(/是)	否
	混淆集法 (confusion－set method) ④	是	是	否/是	否

盖然率高低排列的校正参考清单。为了避免产生新的真词错误,该软件不对真词错误进行自动校正。

词汇语法拼写校对软件由四部分构成,分别对应词汇语法理论中语言的四个级阶(rank scale):词汇处理单元(lexical processing unit)、词组处理单元(group processing unit)、小句处理单元(clause processing unit)和句子处理单元(sentence processing unit)⑤。每个处理单元负责侦别各自语言级阶上的真词错误。每个处理单元内部包含二至三个次处理模块。词汇处理单元由三个次处理模块组成:词汇三元模型(lexical tri－gram model)、概率剖析器(probabilistic parser)、词汇激活单元(lexical activation unit)。词组处理单元和小句处理单元各包含两个部分:类别模型(class model)和概率剖析器。词汇处理单元的类别模型是局域性类别模型(class model (local)),而小句处理单元的是整体性类别模型(class model (global))。句子处理单元也由两部分构成:混淆集模型(confusion－set model)和概率剖析器。

4　词汇语法拼写校对软件的运作机制

由于篇幅所限,本文仅以小句处理单元为例说明词汇语法拼写校对软件中各个部分的运作机制。小句处理单元的工作目标是侦别整体性句法层的真词错误。系统功能语法学派把小句视为基本语法单位。在词汇语法理论体系中,语素构成词,词构成词组、短语,词组、短语构成小句,小句构成句(见图 1)。

由图 1 可知,在小句处理单元,词组是分析的基本单位。小句中的词通常可以被划分成几

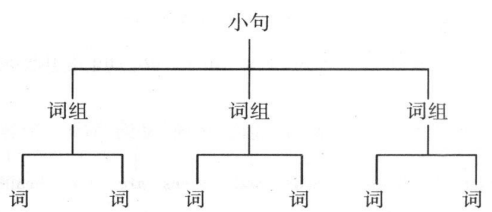

图 1　小句构成

组成为小句的不同成分,每组中有一个或一个以上的词,可以从概念功能的角度界定小句的边界。概念功能是指人们用语言来谈论他们对世界的经验(包括内心世界),用语言来描述周围所发生的事件或情形。世界是由带有属性(形容词)的事物(名词)参与的事件(动词)组成的,事件有其发生的时间、地点、方式等背景(状语)。在功能语法中,小句通过及物性表述这些概念。及物性包括三个成分,即参与者(participant)、过程(process)和环境(circumstance)。过程是及物性系统中的核心成分,它与参与者和环境呈现卫星状的辐射关系,一个过程可以关系到一个或一个以上的参与者和环境。参与者、过程和环境一般由词汇语法中的名词词组、动词词组和副词词组/介词短语体现(参见 Halliday, 1985, 1992, 1994; Thompson, 1996/2000: 13－25, 76－115)(请看表 2)。

因此,在小句处理单元,一切分析都是围绕着过程,即动词词组展开的。小句的界限是通过界定过程－参与者－环境这样的程序来确定的。首先定位句子中的动词词组,然后分析其前面的词汇流和后面的词汇流,界定小句的界限。与动词词组相邻的名词词组将是参与者,

表 2 对小句结构的分析

结构 \ 小句	John	slowly	unlocked	the front door
功能	参与者	环境	过程	参与者
词组	名词	副词	动词	名词词组

副词词组则是环境。这一过程是由小句处理单元的整体性类别模型完成的。最后由概率剖析器在小句内部寻找潜在错误。下面的例子可以具体显示小句处理单元的运作机制。

That herd of cows and calves *are* the healthiest the farm has had in some time. (are 应为 is)

该语句以词汇流的形式进入词汇处理单元,每个单词都被标记上类别(class)(传统语法中称为词性(Halliday, 1994/2000: 28)),并且突显三种词类:名词、动词和副词(见图 2)。

图 2 词类分析

经过词汇处理单元的运作后,该语句进入词组处理单元。词组(group)是词的扩展,主要有三大类:名词词组、动词词组和副词词组(Halliday, 1994/2000: 214),分别体现小句中的不同功能成分。每类词组都有一个中心语,名词词组的中心语通常是事物,由名词充当;动词词组是事件,由(实义)动词体现;副词词组的中心语一般通过副词来体现。词汇处理单元首先识别词组中的中心语,然后根据每个词类的修饰关系确定每个词组的边界。例如:

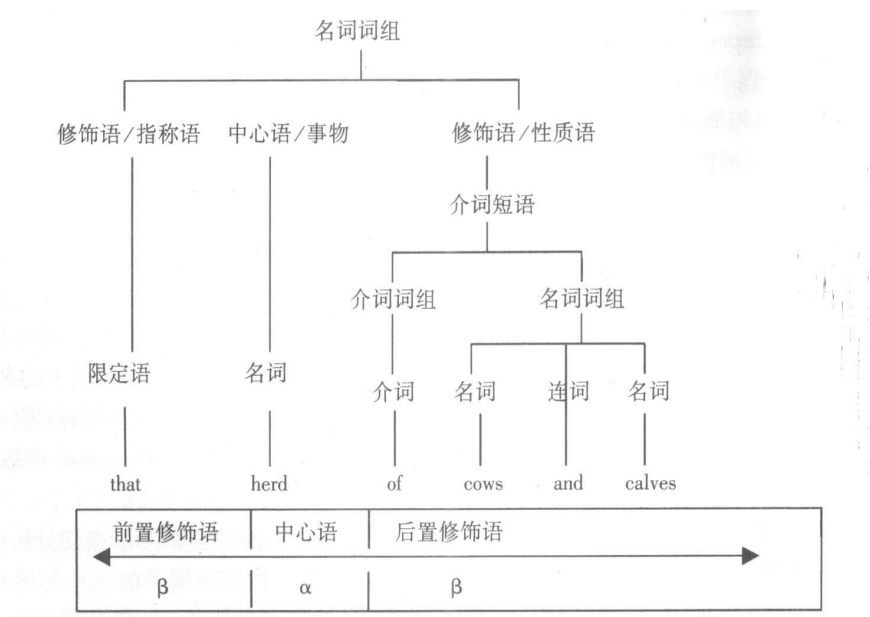

图 3 识别名词词组

在经过词组处理单元的运作后,该输入语句以词组流的形式进入小句处理单元。每个词组的类别标记为:

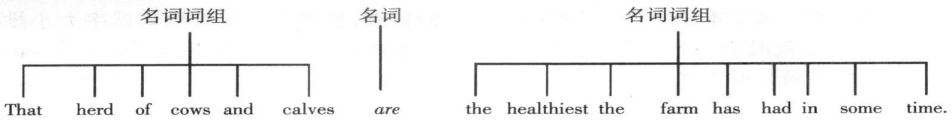

图 4　进入小句处理单元的词组流

该小句中的动词词组只有一个,即 are,它就成为整个小句的中心,体现的是关系过程。然后小句处理单元完成了工作的第一步:定位动词词组。第二步,小句处理单元将要界定整个小句的界限。以动词词组 are 为中心向左看,是名词词组 that herd of cows and calves,它在动词词组的正前方,因此是该关系过程的一个参与者,是被识别者/标示(identified /token)。在动词词组的右侧是另外一个名词词组,它是过程的另一个参与者,是识别者/价值(identifier /value)。至此小句的边界已经界定完毕。该小句中有一个过程、两个参与者。

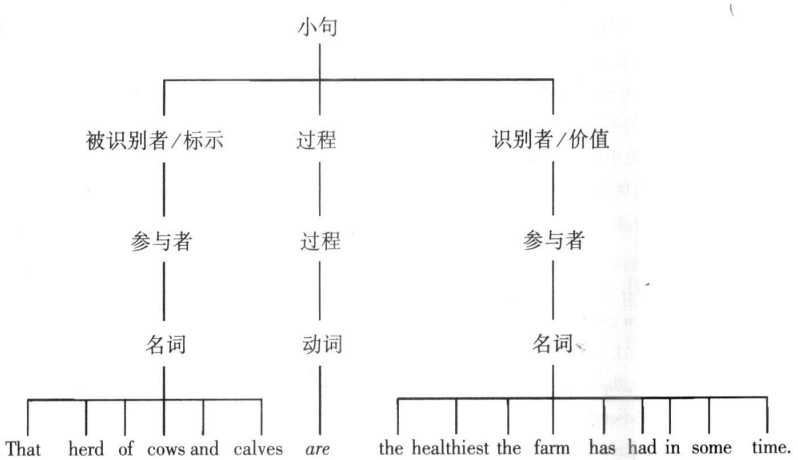

图 5　识别小句边界

小句处理单元工作的第三步就是寻找潜在的违背词汇语法规则的错误。在被识别者/标示和过程之间就出现了不一致。标示的中心语是由 herd 体现的,是第三人称单数形式。为了达到主谓一致,需要动词词组也是第三人称单数形式,而此小句中的动词词组是复数形式。产生这种矛盾有三种可能:一是在动词词组前还有一个标示,二是现有的标示应该被拆分为两个或两个以上部分,显然这两种可能都不存在,第三种可能是动词词组出现了问题。为了达成主谓一致,应该把动词由第三人称复数形式变成单数形式。小句处理单元运作的结果是提供了一个修改建议:由 is 替换 are。

在整个词汇语法拼写校对软件中,各个处理单元是循环运作的。首先,输入文本以词汇流的形式进入词汇处理单元。词汇激活单元包含一个短时内存,可以记忆刚刚被处理过的一部分词。词汇激活单元将这个存储结果和词汇三元模型得出的条件几率(conditional probabilities)相结合,把文本语境因素考虑进来,对校正参考进行排序。其次,词汇处理单元的输出结果被依次送到词组处理单元、小句处理单元和句子处理单元。这些处理单元将侦别各个语言级上的真词错误。最后,句子处理单元的输出结果再次被送到词汇处理单元,将输入文本的前文与后文结合,分析是否存在新的真词错误,并对校正参考进行重新排序。该过程一直循环下去,直到所有的处理单元都同意校正参考的第一建议为止。

5　结论

该词汇语法拼写校对模型的一大创新之处在于以系统功能语言学理论为指导,在构建软件模型的时候融入了词汇语法知识。另外,该模型在进行真词错误纠错工作的时候充分考虑

了文本的语境知识。该文本语境知识不仅仅建立在某词汇前面出现的词汇流上,还包括其后出现的(右侧)词汇流。输入文本的词汇流从头到尾通过词汇语法拼写校对软件中的四个处理单元,任何语言层次上任何类型的潜在真词错误都可以得到纠察。另外,因为输入文本的词汇流在词汇语法拼写校对软件中经过至少两次的检查,最终产生的按照盖然率大小排列的校正参考清单将显示出最适合输入文本的校正参考。在本质上本文解决了一个实际问题,是系统功能语法理论在自然语言生成领域的一次尝试。但由于经济和技术条件的限制,词汇语法拼写校对软件模型尚有待实验检测。

注释:

① CRITIQUE(Heidorn et al., 1982)是由 IBM 公司开发的,前称为 EPISTLE,是根据英文语法规则以单个句子为单位进行纠错的校对软件。

② 该软件由兰开斯特(Lancaster)大学创制,通过分析句子中的每个单词的词性寻找与语法规则相违背的错误,详见 Marshall(1983)和 Garside et al.(1987)。

③ IBM 公司开发的另外一套拼写校对软件,特点在于依托一个两万字的词库,基于期望规则,根据词汇之间组合概率的高低进行纠错。对于该软件的评估参见 Mays et al.(1991)。

④ TriBaySpell(Golding &Schabes, 1996)和 WinSpell(Golding & Roth, 1999)是 Golding 等新近开发的真词错误拼写校对软件,其运作原理是根据预先建立的若干混淆集的不同权重进行校对。

⑤ 该软件中没有包含词汇语法理论中的语素层,因为计算机拼写校对软件属于自然语言处理范畴,其处理的对象是输入词汇流,已经是词汇级阶了,因此语素处理单元可以省略。但如果软件属于自然语言生成领域,如 OCR,则需要增加语素处理单元。

参考文献:

[1] Baddeley, J. A Spelling Checker[J]. *Communications of ACM*, 1980, (5): 456—462.

[2] Eastman, C.M. &D.S. McLean. On the Need for Parsing Ill—formed Input[J]. *Computational Linguistics*, 1981, (4): 257.

[3] Young, C.W., C.M. Eastman & R.L. Oakman. An Analysis of Ill—formed Input in Natural Language Queries to Document Retrieval Systems [J]. *Information Processing and Management*, 1991, (6): 615—622.

[4] Peterson, James L. A Note on Undetected Typing Errors[J]. *Communications of the ACM*, 1986, (7): 633—637.

[5] Grudin, J. Error Patterns in Skilled and Novice Transcription Typing[A]. In W. E. Copper(ed.) *Cognitive Aspects of Skilled Typewriting*[C]. New York: Springer—Verlag, 1983.

[6] Yarowsky, D. Decision List for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French[A]. Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, 1994.

[7] Golding, Andrew R. & Dan Roth. Applying Winnow to Context—sensitive Spelling Correction [A]. Proceedings of the 13th ICML, Bari, Italy, 1996, 182—190.

[8] Kukich, Karen. Techniques for Automatically Correcting Words in Text[J]. *ACM Computing Surveys*, 1992, (4): 377—439.

[9] Heidorn, G. E. et al. The EPISTLE Text—critiquing System[J]. *IBM Systems Journal*, 1982, (3): 305—326.

[10] Marshall, Ian. Choice of Grammatical Word—class without Global Syntactic Analysis: Tagging Words in the LOB Corpus[J]. *Computers and the Humanities*, 1983, (17): 139—150.

[11] Garside, Roger, Geoffrey Leech &Geoffrey Sampson. *The Computational Analysis of English: A Corpus—based Approach*[M]. London: Longman, 1987.

[12] Mays, Eric, Fred J. Damerau & Robert L. Mercer. Context—based Spelling Correction[J]. *Information Processing and Management*, 1991, (5): 517—522.

[13] Golding, Andrew R. & Yves Schabes. Combining Trigram—based and Feature—based Methods for Context—sensitive Spelling Correction[A]. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA, 1996, 71—78.

[14] Bloor, Thomas & Meriel Bloor. *The Functional Analysis of English: A Hallidayan Approach*[M]. London/Beijing: Arnold/FLTRP, 1995/2001.

[15] Halliday, M.A.K. *Short Introduction to Functional Grammar*[M]. London: Arnold, 1985.

[16] Halliday, M.A.K. Language as System and Language as Instance: The Corpus as a Theoretical Construct[A]. In Jan Svartvik(ed.) *Directions in Corpus Linguistics*[C]. Berlin: Mouton de Gruyter, 1992, 61—77.

[17] Halliday, M.A.K. *An Introduction to Functional Grammar*[M]. London/Beijing: Arnold/FLTRP, 1994/2000.

[18] Thompson, Geoff. *Introducing Functional Grammar*[M]. London/Beijing: Arnold/FLTRP, 1996/2000.

[19] 张磊. 中文文本自动校对[J]. 语言文字应用, 2001, (1): 19—26.