**文 献 综 述**

课 题 名 称 基于文本挖掘的汉语语法检查

院　 系 信息科学与技术学院

专 业 班 计算机技术

姓　 名　 余冠源

评 分

指 导 教 师 张治国

1. 概述
   1. 选题背景
      1. 问题的提出

随着互联网计算机的兴起，越来越多的文书工作已经离不开计算机。而随着计算机的普及，越来越少人会直接在纸上书写汉字。传统的笔录汉字已经慢慢迁移到计算机上进行。当需要检查一篇文章，是否有错别字或者语法错误的时候，传统的做法是需要审核者肉眼一字一句的查看，人为的判断文章是否存在错误词句，或者标点符号错误。而随着文书工作转移到计算机，文章审阅方法仍然离不开传统的人为判断。

近年来，随着自然语言处理（NLP）和文本挖掘算法的兴起，使得计算机处理文本的能力再不断增强。同时，融合传统的数据挖掘方法，使得通过计算机做初步的文章语法检查称为可能。计算机对一篇文章做语法分析，主要从语法、结构和内容三方面入手，其中又涉及语言的复杂性，表达的恰当性，语言的流利性等。利用计算机做语法分析主要有两种方法，基于规则的方法和基于统计的方法。基于规则的方法应用自然语言处理技术对文本进行语法分析和语义分析，达到校对错误的目的。另外，基于统计模型你的方法是通过使用语料库和统计语言模型对待校对文本进行统计计算，并设定一个正确文本的阈值，通过计算的结果与阈值比较，从而判断原文是否存在错误。

* + 1. 国内外研究现状

文本挖掘是描述一系列旨在为商业智能，数据分析统计提供数据来源的，结合语义，统计和机器学习的建模方法。劳动密集型人工文本挖掘方法始于20世纪80年代。文本挖掘交织着信息获取，数据挖掘，机器学习，统计和计算机语义等领域，是一个跨多个领域的数据处理方法。因为大部分的数据（普遍认为超过80%）[[1]](#引文1)是通过文本保存，所以文本处理被认为由相当高的商业价值。

同时，随着计算机计算能力的摩尔定律式的提高，使得计算机处理文本的能力大大增强。文本挖掘算法开始关注语料库语言学，之后语言模型也不再是基于人工书写规则，而是更加关注统计模型。结果的判断也是基于概率统计结果。如今的文本挖掘算法，可以建立再大量的训练数据集上，统计得出模型的参数，各个词汇的出现频率和词与词之间的关联概率。

目前，国外的研究方法基本分为两大类，基于规则的语法检查和基于统计的语法检查[[2]](#引文2)。基于规则的语法检查基本思路是构建语法规则库，通过与规则库进行匹配来判断有没有语法错误。基于规则的方法研究起步比较早，主要是通过模式匹配的方法，而近年来，又发展出基于负实例的语法检查算法。另外基于统计的语法检查则基于N-GRAM语法模型。N-GRAM模型构建简单直接。目前主要的构建思路为：将整个句子视为一个字符串，处理句子得到其中的N元语法后，根据乘法原理，求出整个字符串的先验概率，通过比较计算出先验概率和一个阈值来判断是否又语法错误。或者，将句子视为一个字符串，处理句子得出N-GRAM语法，将每一个N-GRAM语法的频次和一个阈值做比较，从而判断是否又语法错误。

在国内，中文语法错误检查基本也是基于统计方法和基于规则的方法。这些方法对语法校对都存在着一些不足，例如统计方法使用词的相邻矩阵N元模型，只能反应局部的语法限制，而不能包含长距离的语法限制。而基于规则的方法则是自底向上或自顶向下的分析方法，不能发现特定词语的搭配错误，同时时间复杂度也相当的高。另外，也有一些研究人员针对中文语法的特点，使用模式匹配和句型分析相结合的语法检查方法[3]。

以下将选取一些比较常见的语法检查方法，并做简要阐述：

* + - 1. 基于PSG的句法分析

另外近些年来，还发展了基于PSG的句法分析[[4]](#引文4)。基于规则的自动语法检测的基本原理是利用语法规则对句子进行分析，合乎语法的句子会生成完整的树形结构。这种利用句法分析检测错误的方法源于PSG。PSG中重写规则由非终结符和终结符构成。非终结符指短语类别如名词短语NP、动词短语VP和单词词性如动词Verb、名词N等，可以置于规则X->Y的两侧；终结符指具体的词汇如man、ball等，只能置于规则的右侧。

* + - 1. 基于特征结构的句法分析

特征结构是对PSG语法规则的拓展，目的是限制其过于宽泛的生成能力。特征结构一般用“属性-值”的方式表述，属性由特定的语言学信息如人称、数、格和次范畴等构成，每一类属性的属性取值范围形成一个集合，如数={单数 复数}。同时，特征结构的本质是一个递归表达式，各成分之间的属性可以互相嵌套。

* + - 1. 基于词汇的句法分析

链语法是以词汇中心主义为基础进行句法切分的代表。链语法由词典和算法两部分组成，词典中包括人工编写的基于词汇的句法搭配方式；算法根据每个此条的搭配方式对句子进行切分。链语法词典由词条和链接子表达式构成，链接子表达式中包括句法关系名称；冠词-名词（D）、谓词-宾语（O）和主语谓语（S）；两种链接方向：+ 表示朝右；- 表示朝左；两个逻辑操作符：合取& 和析取 or。

* + - 1. EasyEnglish

EasyEnglish是由IBM为英语学习者开发的语法检查器，它是基于“英语槽语法”（English Slot Grammar）开发的，是通过搜索用网络图表示句法的树来发现错误[[5]](#引文5)。语法错误被形式化为模式，以便与句法树进行匹配。EasyEnglish是基于规则的方法实现的，但是它没有公开的发行版本。

* + - 1. 啄木鸟系统

啄木鸟系统是基于这么一个事实，绝大多数包含错误语法的句子在分词之后，会出现落单的单字词。对于落单的单字词，根据其词频和前后两个汉字的连续强度给该单字词打分，最后将得分与一个预先设定的阈值作比较，判断该落单字词是否为错误字词。

* + - 1. 张照煌的近似字集替换法

张照煌首先根据四种汉字的相似类型：同音或近音，字形相近，字义相近，输入码相近，生成各个汉字C的相似字集cfs(C)。这与英文中的每个混淆集概念类似。该方法对句子中各个汉字使用近似字集中的汉字依次替换，得到

| cfs(C1) |×| cfs(C２) |×．．．×| cfs(Cｎ) |

个候选字符串，然后用词间字二元模型和词性二元模型对各个字串进行评分，并考虑非原字扣分，最后选出得分最高的字符串，其中的汉字与原句子中对应的汉字不同的地方被认为有错。

* 1. 相关技术背景
     1. 马尔可夫模型

在概率论中，马尔可夫模型是一个随机模型，用于对随机变化的系统上建模，而这个随机系统是基于一种假设，未来的状态只与该状态前的多个状态有关。马尔可夫模型常用于在如下表1-1的场景当中：

表格 1-1 马尔可夫模型应用场景

|  |  |  |
| --- | --- | --- |
|  | 系统状态可以完全监测 | 系统状态只可以部分被监测 |
| 系统是自治的 | 马尔可夫链 | 隐藏马尔可夫模型 |
| 系统是可控制的 | 马尔可夫决策过程 | 部分可监测的马尔可夫决策过程 |

最简单的马尔可夫模型是马尔可夫链。马尔可夫链通过随时间变化的随机变量对系统状态进行描述。马尔可夫假设变量的分布只与前几个状态相关。马尔可夫链的一个典型例子是，马尔可夫蒙特卡罗方法（MCMC）。在统计学中，马尔可夫蒙特卡罗方法是对概率分布进行采样的一类算法，该算法基于构建马尔可夫链，当所求解问题是某种随机事件出现的概率，或者是某个随机变量的期望值时，通过某种“实验”的方法，以这种事件出现的频率估计这一随机事件的概率，或者得到这个随机变量的某些数字特征，并将其作为问题的解。

隐马尔可夫模型是一种马尔可夫链，但是该系统中，只有部分状态可以被观察到。换而言之，观察到的现象是依赖于系统的状态，但是我们却不能从观察到的现象准确的预测系统的状态。常见的隐马尔可夫模型算法有几种，例如给定一个系统观察序列，维特比（Viterbi）算法可以计算出哪个系统状态会有最大的出现概率。另外，前向算法可以用于计算观察序列的概率，前向-后向算法[[6]](#引文6)则是用于对隐马尔可夫模型的参数进行估计，然后通过给定的数据评估这些参数的价值并减少它们所引起的错误来重新修订参数。

马儿可夫决策过程是包含当前状态和一个决策向量的马儿可夫链。马儿可夫决策过程一般用于计算在给定的规则的上下文中，最大化期望的回报。

部分可观察的马儿可夫决策（POMDPs）过程是描述一个系统，只有某些状态能被观察到的这么一个马儿可夫决策过程。POMDPs被普遍认为是NP完全问题。通常在实际应用当中，会使用近似算法和技术。

* + 1. N-GRAM模型

基于统计方法的建模方法主要是基于N-GRAM语法模型。N-GRAM语法模型构建简单，基于大型语料库的训练即可得到。目前，N-GRAM语法模型在多个自然语言处理领域使用。N-GRAM模型是一种基于概率语法的模型，用于预测序列中下一个元素。N-GRAM模型实际上是N-1序列的马尔可夫模型。N-GRAM模型有两个核心的优点，分别是实现的相对简单性和容易扩展的特性（在时间和空间的复杂度之间平衡的前提下，只需简单增加N的值，实现存储更多的上下文信息）。

N-GRAM模型的概念来源于Claude Shannon在信息论上的工作[[7]](#引文7)。Shannon提出了一个问题：给定一个字母序列，例如，序列 “for ex”，那么哪一个字母会有最大的可能性出现在序列的下一个位置？从训练数据上，我们可以从一个N长度的序列中，计算字母的概率的分布，而每一个字母的概率的和为1。更确切的说，N-GRAM模型是基于序列Xi-(n-1) ,…, Xi-1预测Xi的值。按概率论的说法，就是P(Xi| Xi-(n-1) ,…, Xi-1)。在自然语言建模当中，通常会加上词语独立性假设，假设每一个词语仅仅依赖于它前面的n-1个词语。这个马尔可夫模型旨在用于迫近真实的自然语言。这个假设非常关键，因为这个假设大大简化了问题的复杂度。

由于N-GRAM模型拥有以上的这些优点，所以N-GRAM模型被广泛应用于自然语言处理，语音识别等领域。虽然N-GRAM模型有着不少优点，但是N-GRAM模型也有一些缺点而常常饱受诟病。首先，N-GRAM模型因为基于的n-1个词语的上下文依赖假设，所以该模型不能很好的处理噪声。

1. 相关技术分析

本章节主要介绍本文所用到的相关技术和系统架构，包括N-GRAM模型，HMM隐马尔可夫模型。

* 1. 系统实现

本论文将采用N-GRAM模型与隐马尔可夫模型相结合的办法实现中文汉字的语法检查。中文语法检查首先必须建立一个语料库，供系统的模型训练。然后以N-GRAM和隐马尔可夫算法建立模型，再对输入文章进行概率计算，从而得出错误概率。系统的模型，如图2-1所示：

词典

错误句法语料库

文章切分为句

待检查文章

句子切分为词

词性标注

训练模型

计算每个词的错误概率

错误词语标记

N-GRAM模型

HMM 隐马尔可夫模型

结巴分词

中科院汉语词法分析系统 ICTCLAS)

图 2-1 系统实现模型

* + 1. 语料库的建立

系统利用一个已有的大规模的已进行了分词和词性标注的现代汉语语料库。该语料库包含了1998年1月至12月的《人民日报》中466万字的文章。通过语料库进行统计，得到汉语文本检查所需的语言规律和知识，包括字词概率表、字词二元同现概率表、词性概率表、词性二元同现概率表，并将这些概率表作为中文文本检查的基础数据。

* + 1. 句子分词和标注

对于待检查的文章，首先就是对文章进行分句。然后在分句得出的句子基础上，再进行分词。目前，中文分词已有相当成熟的模型和算法可以使用，这些算法基本是基于隐马尔可夫模型。例如开源的软件有，结巴分词。还有中国科学院计算技术研究所研发出来的汉语词法分析系统ICTCLAS。这些分词系统都有相当高的分词效率和精确率。因为本论文重点研究的方向是语法错误的检查，所以在分词的环节上，使用已有并且成熟的技术。

* + 1. 词语相关性计算

在完成了分词之后，重点就是在分词的基础上计算每个词语与上下文的关联程度。在这里，本论文使用了两个方向词语的概率分布计算。在上一个步骤中，词语已经被标注了词性，根据语料库计算的词性上下分分布概率，可以提供一个阈值来判定词性是否合理正确。举例说明，如图2-2：

我爱背景天安门

我/D 爱/V 背景/N 天安门/N

切词与词性标记

P(D|\*,\*)

P(V|\*,D)

P(N|D,V)

P(N|V,N)

词性的概率计算

P(我|\*,\*)

P(爱|\*,我)

P(背景|我,爱)

P(天安门|爱,背景)

阈值检查

标记错词概率

词语的概率计算

图2-2 模型计算流程样例

系统输入是这么一个句子：“我爱背景天安门”，分词和词性标注后，得到如下的词与词性的二元组合（假设使用三元马儿可夫模型）：

（\*，\*）（\*，\*）（我，D），（爱，V），（背景，N），（天安门，N）

在句子的开头，加入了两个假定的句法开头标记（\*），这是为了让每个字词都可以计算三元马儿可夫概率分布，包括开头的词（我）。得到二元组合后，首先计算词性的三元马儿可夫概率：

P11(D|\*, \*)，P12 (V|\*, D)，P13(N|D, V)，P14(N|V, N)

然后，就需要计算每个词的三元马儿可夫概率：

P21(我|\*,\*)，P22(爱|\*, 我)，P23(背景|我, 爱)，P24(天安门|爱, 背景)

在计算完词性和词语的概率后，再做一个联合的基于权重的加法：

E(Ci) = W1i×P1i(Ci|Ci-2, Ci-1) + W2i×P2i(Ci|Ci-2, Ci-1)

通过以上公式计算得出的E(Ci)即为Ci词正确的概率，E值越小，则该词被认定为错误的概率越大。并 根据E值的大小，标注在原文当中，提示该词有多大的可能性是错误的词语。

参考文献

[1] Breakthrough Analysis. Unstructured Data and the 80 Percent Rule. 2008-08-01，<http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/>

[2] 葛昊. 基于N-gram语法检查系统的改进和实现[D]:硕士. 中国科学技术大学. 2014(11-12)

[3] 龚小谨,罗振声,骆卫华. 中文文本自动校对中的语法错误检查[J]. 计算机工程与应用. 2003(01)

[4] 刘磊. 面向自动语法检查的依存规则研究[D]:硕士. 北京外国语大学. 2014(28-38)

[5] 刘雷. 英语作文智能批改中语法检查的研究与实现[D]:硕士. 北京邮电大学. 2012(10-11)

[6] Forward-backward algorithm. http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html\_dev/forward\_backward/s1\_pg1.html

[7] Wikipedia. <https://en.wikipedia.org/wiki/N-gram>